Attachment preferences and corpus frequencies in PP attachment ambiguities: Evidence from Greek

Kalliopi Katsika
Aristotle University of Thessaloniki
kkatsika@enl.auth.gr

Abstract
The present study investigates the extent to which frequency affects Greek native speakers’ attachment preferences for ambiguous Prepositional Phrase structures in Greek. Analyses of samples from a written and a spoken Greek corpus were conducted in order to examine the frequency patterns of PP attachment in Greek. The results of the corpus analyses were contrasted to the findings of an on-line self paced reading task on the processing of PP attachment. The results indicate that corpus frequencies correspond to parsing decisions on a coarse-grained (syntactic) level, and to a lesser extent on a fine-grained (lexical) level of analysis.

Keywords: attachment, corpus, frequency, prepositions, Prepositional Phrases, sentence comprehension, structural ambiguity

1. Introduction
It has recently been suggested in the psycholinguistic literature that language processing may be tightly connected to prior linguistic experience (e.g. Cuetos & Mitchell 1988; Desmet, De Baecke, Drieghe, Brysbaert & Vonk 2006; Desmet & Gibson 2003; Gibson & Schütze 1999; Gibson, Schütze & Salomon 1996; Igoa, Carreiras & Meseguer 1998; MacDonald, Pearlmuter & Seidenberg 1994; Mitchell, Cuetos, Corley & Brysbaert 1995). Indicative of this interest in frequency based parsing is the appearance – in the last two decades – of statistically based natural language processing approaches to knowledge representation, processing and learning. These approaches include probabilistic models of sentence parsing (e.g. Crocker & Brants 2000; Jurafsky 1996; Sturt, Costa, Lombardo & Frasconi 2003) and connectionist models of sentence processing and production (e.g. Altmann 2002; Rohde 2002; Tabor, Juliano & Tanenhaus 1997) that are capable of learning grammatical patterns based on previous experience. In addition, an increasing number of psycholinguistic studies have started from corpus analyses in order to examine whether the most frequent structure is also the easiest to process in online parsing of syntactically ambiguous sentences (e.g. Desmet, Brysbaert, De Baecke 2002; Desmet & Gibson 2003; Gibson & Schütze 1999; Gibson, Schütze & Salomon 1996; Mak, Vonk & Schriefers 2002; Mitchell & Brysbaert 1998).

The aim of the present study is to investigate the frequency of prepositional phrase (PP) attachment in Greek and to contrast the frequency results with native Greek speakers’ online parsing preferences for ambiguous V-NP-PP structures. The comparison between these two types of linguistic evidence was conducted in order to

---

1 This study is part of the 03ED375 research project, implemented within the framework of the “Reinforcement Programme of Human Research Manpower” (PENED) and co-financed by National and Community Funds (25% from the Greek Ministry of Development – General Secretariat of Research and Technology and 75% from E.U. – European Social Fund).
examine the assumptions of the Tuning Hypothesis (Cuetos, Mitchell, & Corley 1996) which directly links parsing preferences to structural frequency, as it assumes that the resolution of ambiguous structures is determined by the reader’s or listener’s prior experience / exposure to ambiguities of the same kind. The extent to which the Tuning Hypothesis can actually account for parsing preferences has been under investigation in the psycholinguistic literature.

2. The PP attachment ambiguity
The V-NP-PP ambiguity involves possible attachment of a PP either to the preceding VP or to the preceding NP and can be illustrated in structures such as:

(1) O kataskopos idhe ton andhra me to tileskopio.
    The-NOM spy-NOM saw-PERF.3S the-ACC man-ACC with the-ACC telescope-ACC
    ‘The spy saw the man with the telescope.’

In verb-noun phrase-prepositional phrase (V-NP-PP) sequences such as (1), the ambiguity lies in the possibility of attaching the PP me to tileskopio either to the preceding verb idhe denoting the instrument of the action described by the verb, or to the preceding NP ton andhra, as a modifier of the NP. It has been claimed in the psycholinguistic literature that native speakers’ initial attachment preferences may be based upon different types of information such as syntactic (Frazier 1987; Rayner, Carlson & Frazier 1983), subcategorization (e.g. Cliffton, Speer & Abney 1991), discourse level / lexical information (e.g. MacDonnald, Perlmutter & Seidenberg 1994) or structural frequency (Cuetos et al. 1996).

Spivey-Knowlton & Sedivy (1995) conducted a study in which they examined corpus frequencies and native English speakers’ attachment preferences for temporary ambiguous with-PP structures. Their experimental material – based on Altmann & Steedman’s (1988) stimuli – included sentences such as:

(2a) The fireman smashed down the door with the/a rusty lock. (NP attach-definite)
(2b) The fireman smashed down the door with the/a heavy axe. (VP attach-definite)
(2c) The fireman smashed down a door with the/a rusty lock. (NP attach-indefinite)
(2d) The fireman smashed down a door with the/a heavy axe. (VP attach-indefinite)

The results of two online self-paced reading tasks revealed that participants mean reading times were generally faster in the VP attachment than the NP attachment condition. On the other hand, the corpus analyses showed that definite NPs followed by with were highly biased towards VP attachment whereas indefinite ones were biased towards NP attachment. Finer-grained analyses in which action verbs were analysed separately from psychological/perception predicates showed a correspondence between corpus frequencies and online attachment preferences.

Hindle & Rooth (1993) hand-parsed a sample of 880 randomly selected instances from the 1989 Associated Press News Stories corpus and they observed that NP attachment of PP was more frequent (67%) than VP attachment. These results contradict Spivey-Knowlton & Sedivy’s (1995) corpus analysis which revealed an overall advantage for VP attachment. A possible reason for this discrepancy (as noted by Spivey-Knowlton & Sedivy, 1995) might be that Hindle & Rooth (1993) included a variety of prepositions in their analysis whereas Spivey-Knowlton & Sedivy (1995)
examined attachment frequencies of only with-PPs which could be biased towards instrumental VP attachment.

3. The Present Study
The present study challenges the predictions of the Tuning Theory (Mitchell et al. 1995) by examining PP attachment frequency both at a coarse-grained level (VP vs. NP attachment) and at a more fine-grained level (lexical choice of preposition, definiteness of the prepositional NP complement). More specifically, the aim of this study is to examine which type of PP attachment (VP or NP) is the most frequent in naturally produced sentences in written and spoken register and to investigate the possible effect of lexical factors on the frequency of VP vs. NP attachment, as well as the influence of a language specific grammatical phenomenon (Definiteness Agreement) on PP attachment. The study includes sample analyses of two types of corpora; a written (Institute for Language and Speech Processing corpus) and a spoken language corpus (manually compiled for the purposes of this study).

3.1 ILSP corpus analysis
3.1.1 Method, techniques, materials
The written corpus sample was extracted from the Hellenic National Corpus (ILSP corpus). Two different corpus sets were extracted from the ILSP corpus; a small ‘unrestricted’ set (2,000 sentences) which was extracted on the basis of coarse grained criteria, and a larger ‘restricted’ set (77,744 sentences) whose extraction was based on more fine grained criteria. The main difference between the two sets is that the extraction of the unrestricted sample was based on purely structural criteria (only major category information were inserted in the ILSP corpus engine) whereas the extraction of the restricted corpus sample included searches on the basis of prepositions me, se, apo and ja. The basic rationale behind the extraction and analysis of these two types of corpus samples was to check possible differences between coarse grained and fine grained frequency data.

The extracted V-NP-PP structures were first filtered manually so as to make sure that each sentence actually included a V-NP-PP structure. This filtering process resulted in the deletion of standard expressions, idioms and all types of sentences that did not correspond to the relevant structure. The filtering process resulted in the deletion of 65% of the sentences in the unrestricted set and 58% of the sentences in the restricted corpus set. The remainder of the sentences (697 in the unrestricted set, 32,594 in the restricted set) were analysed further and the results of these analyses are presented in the following section.

3.1.2 Results
The analysis of the 697 sentences of the unrestricted set revealed that PPs attach to the preceding VP more frequently than to the preceding NP (see table 1). This result was supported statistically by a chi-square goodness of fit test which showed a significant difference between VP and NP attachment; $\chi^2 (1) = 36.271, p < .000$. Similarly to the results of the unrestricted corpus set, there was an overall VP attachment advantage in the restricted corpus set ($\chi^2 (3) = 3527.745, p < .001$).

Table 1. PP attachment in the ILSP corpus.

<table>
<thead>
<tr>
<th>Corpus set</th>
<th>Attachment site</th>
<th>Corpus sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N</td>
</tr>
</tbody>
</table>
Further analyses were conducted in the restricted corpus set in order to examine the possible effect of the lexical choice of preposition (me, se, apo, ja). The results of VP and NP attachment per preposition appear in table 2.

Table 2. PP attachment per preposition in the ILSP corpus.

<table>
<thead>
<tr>
<th>Attachment site</th>
<th>ILSP corpus sentences per preposition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>me</td>
</tr>
<tr>
<td>N %</td>
<td></td>
</tr>
<tr>
<td>VP</td>
<td>3,814</td>
</tr>
<tr>
<td>NP</td>
<td>2,121</td>
</tr>
<tr>
<td>Total</td>
<td>5,935</td>
</tr>
</tbody>
</table>

As it can be seen in table 2, the pattern of attachment in sentences with me-, se- and apo-headed PPs was similar to the general pattern of attachment; VP attachment were significantly more frequent than NP attachments (me: $\chi^2 (1) = 482.94, p < .001$, se: $\chi^2 (1) = 4039.49, p < .001$, apo: $\chi^2 (1) = 382.252, p = .000$). In contrast, ja-headed PPs attached significantly more frequently to the preceding NP than to the preceding VP; $\chi^2 (1) = 827.407, p < .001$.

In addition, the analysis of the corpus counts in relation to the definiteness of the NP complement of P allowed for the investigation of whether Definiteness Agreement extends to complex DPs (DP+PP) in the written corpus data (see Stavrou & Tsimpli 2009). The results of this analysis can be seen in table 3.

Table 3. PP attachment per NP2 definiteness and per preposition in the ILSP corpus.

<table>
<thead>
<tr>
<th>Attachment site / NP1-NP2 definiteness</th>
<th>ILSP corpus sentences per preposition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>me</td>
</tr>
<tr>
<td>N %</td>
<td></td>
</tr>
<tr>
<td>VP definite-definite</td>
<td>1,544</td>
</tr>
<tr>
<td>VP definite-indefinite</td>
<td>227</td>
</tr>
<tr>
<td>Total</td>
<td>1,771</td>
</tr>
<tr>
<td>NP definite-definite</td>
<td>1,026</td>
</tr>
<tr>
<td>NP definite-indefinite</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>1,032</td>
</tr>
</tbody>
</table>

As illustrated in table 3, definite-definite structures were more frequent than definite-indefinite ones in both VP and NP attachment throughout all prepositions. More specifically, there was a significant association between definiteness and attachment in preposition me ($\chi^2 (1) = 128.093, p < .001$) and se structures ($\chi^2 (1) = 78.010, p < .001$). This indicates that there were differences between the definiteness percentages in VP and NP attachment, although definite-definite structures were significantly more frequent than definite-indefinite in VP (me: $\chi^2 (1) = 979.384, p < .001$, se: $\chi^2 (1) = 6992.988, p < .001$) and NP attachment (me: $\chi^2 (1) = 1008.14, p < .001$, se: $\chi^2 (1) =$...
The differences in definiteness percentages stem from the fact that NP attachment definite-indefinite structures were the least frequent ones for both preposition me (0.6%) and preposition se (1.8%) sentences. On the other hand, there was no significant association between definiteness and attachment in apo- and ja-headed PP structures. This means that the distribution of definiteness in VP and NP attachment was similar in the case of preposition apo and ja sentences.

These results provide an indication that the lexical content of preposition plays a role in the licensing of Definiteness Agreement in complex DPs. Although definite-definite structures NP attachment structures are strikingly more frequent than definite-indefinite ones throughout all prepositions, it is only in the case of prepositions me and se that the definite-indefinite NP attachment structures are significantly fewer than the definite-indefinite VP attachment structures.

3.2 Spoken corpus analysis
3.2.1 Method, techniques, materials

The main purpose of the spoken language corpus study was to compile a plausible sample of standard Modern Greek. In order to achieve this goal, 37 Greek TV shows of various types were transcribed into a manually compiled corpus of 347,107 words. All recordings were TV broadcast data and were obtained online via web pages that allow free watching of a variety of broadcasts (www.greek-movies.com and www.livemovies.gr). Each transcription was saved as a separate text document in which searches with concordance software (MonoConc Pro) were conducted in order to find the relevant V-NP-PP structures. Each search was made on the basis of prepositions me, se, apo and ja. Thus, the spoken language corpus was restricted in the sense that it only included instances of PPs headed by prepositions me, se, apo and ja but it was not restricted in terms of the presence of a determiner on the NP complement of the verb.

3.2.2 Results

A total of 1182 V-NP-PP structures were extracted from the spoken corpus files. After an initial analysis of the data, 136 sentences were excluded from further analyses because they were characterised as Standard Greek Expressions (e.g. Είδε το Χάρο με τα μάτια της) and Light Verbs (e.g. Έχω μια συμπάθεια στον Ολυμπιακό Βόλου...). Further analyses were thus conducted in the remaining 1046 spoken corpus sentences. The results of PP attachment frequency per preposition appear in table 4.

<table>
<thead>
<tr>
<th>Attachment site</th>
<th>Spoken corpus sentences per preposition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>me</td>
</tr>
<tr>
<td>VP</td>
<td>N</td>
</tr>
<tr>
<td>VP</td>
<td>118</td>
</tr>
<tr>
<td>NP</td>
<td>83</td>
</tr>
<tr>
<td>Total</td>
<td>201</td>
</tr>
</tbody>
</table>

Overall, VP attachments were significantly more frequent than NP attachments; $\chi^2 (1) = 157.587, p < .001$. In addition, a multidimensional chi-square test indicated a significant interaction between attachment and preposition ($\chi^2 (3) = 150.489, p < .001$) and this means that VP and NP attachment varied across prepositions. Indeed, VP attachment was significantly more frequent than NP attachment in prepositions me ($\chi^2 (1) = 6.095, p = .014$), apo ($\chi^2 (1) = 8.000, p = .005$) and se ($\chi^2 (1) = 260.963, p <
.001). On the other hand, preposition ja PPs attached significantly more frequently to the preceding NP than to the preceding VP; $\chi^2(1) = 10.347, p = .001^2$.

3.3 Discussion of corpus findings
3.3.1. PP attachment
The analysis of the ILSP corpus and spoken language corpus revealed a general VP attachment advantage on a coarse-grained level of analysis. This result is incompatible with the main finding of Hindle & Rooth’s (1993) English corpus data in which NP attached prepositional phrases were more frequent than VP attached PPs. The present data showed a precedence of NP attachment only on a fine-grained level of analysis, in the case of preposition ja. The pattern of attachment of ja-headed PP was the opposite of the pattern of attachment of prepositions me, se and apo which attached more frequently to the preceding VP than to the preceding NP. More specifically, me- and apo-PPs showed a similar pattern of attachment whereas se-PPs exhibited significantly higher VP attachment percentages than prepositions me and apo. This difference is most probably linked the degree of lexical information that prepositions me, se, apo and ja carry; the fact that preposition se has very little semantic content results in the strong association of se-PPs with the preceding VP whereas the rich lexical content of preposition ja makes ja-PPs favour NP attachment. Prepositions me and apo are found somewhere in the middle as their attachment pattern may be subject to discourse and lexical factors (e.g. definiteness, animacy).

3.3.2 Definiteness Agreement
Definiteness Agreement was predicted to affect the frequency of NP attachment in the definite-definite and definite-indefinite conditions. Assuming that the possibility of postulating multiple definite determiners in Greek nominals extends to determiners occurring in complex DPs (DP+PP) (see Stavrou & Tsimpli 2009 for complex subject DPs), definite-definite NP attachments were expected to be more frequent/preferred than definite-indefinite NP attachments. The results of the ILSP corpus data indicated that the percentage of definite-indefinite NP attachments was significantly smaller than the percentage of definite-indefinite VP attachment in structures with preposition me and se. These results could be an indication that Definiteness Agreement has an influence on complex object DPs depending on the choice of preposition which probably indicates differences in the P’s lexicality (see e.g. Terzi 2007; Papadopoulou et al. 2007).

4. The role of frequency in the parsing of ambiguous PP structures
The analysis of the ILSP and spoken language corpus provided an interesting insight to the patterns of PP attachment on coarse-grained and fine-grained levels of analysis. In addition, the investigation of PP attachment in two different registers showed very little variation between the written and spoken text sample. It can thus be safely argued– at least for the type of structures studied here – that structural frequency patterns do not generally differ between written and spoken corpora. What then remains to be seen is the extent to which corpus frequencies are reflected on Greek native speakers’ psycholinguistic preferences for temporarily ambiguous PP structures.

---

2 An analysis in order to examine the possible effect of Definiteness Agreement was not conducted in the spoken language corpus because of the very limited number of data (no definite-indefinite structures in prepositions me, apo and ja and 1 sentence in preposition se).
4.1 Online processing of PP attachment
Native Greek speakers’ initial attachment preferences for temporarily ambiguous V-NP-PP structures were investigated in an online self-paced reading task reported in Katsika (2009). Forty seven adult native speakers of Greek participated in this study. The experimental stimuli included sentences in which the lexical choice of the preposition (me ‘with’, se ‘in, into’, ja ‘for’, apo ‘from, by’) and definiteness (definite vs. indefinite NP object of P) were manipulated. The experimental sentences had the following form:

(3a) O kipuros ekopse to kladhi me to/ena maxeri. (VP attachment def/indef)
    The-NOM gardener-NOM cut-PERF.3S the-ACC branch-ACC with the/a-ACC knife-ACC
    ‘The gardener cut the branch with the knife.’

(3b) O kipuros ekopse to kladhi me to/ena luludhi. (NP Attachment def/indef)
    The-NOM gardener-NOM cut-PERF.3S the-ACC branch-ACC with the/a-ACC flower-ACC
    ‘The gardener cut the branch with the flower.’

The online self-paced reading task showed that Greek native speakers read VP attachments faster than NP attachments. The lexical choice of the preposition had a significant effect on participants reading times; VP attachments were preferred when sentences included PPs headed by prepositions me, apo and ja, while there was no attachment site preference for sentences with se-headed PPs. In addition, Definiteness Agreement structures were preferred in sentences with prepositions se and me (but not with apo and ja). It should also be noted that the significant difference between NP attachment definite-definite and definite-indefinite sentences found only in the critical PP segment provides additional evidence for the localised nature of the grammatical phenomenon of Definiteness Agreement.

4.2 Parsing data vs. frequency data
On a coarse-grained level of analysis, the Tuning Hypothesis (Mitchell et al. 1995) predicts that readers’ parsing preferences are expected to be ‘tuned’ to the most frequent relevant structures in a given language and thus corpus data should be compatible with readers’ online parsing decisions. The analysis of the corpus data in the present study provided a measurement of the most frequent patterns of PP attachment in Greek. Under a strict version of the Tuning Hypothesis (adopted by Don Mitchell in Mitchell et al. 1995), only coarse-grained level information should be available to the parsing in the initial processing of a (temporary) ambiguous string. Even though Mitchell et al. (1995) presume that the more detailed the analysis the more the chances for a correspondence between frequency data and parsing preferences, they dismiss the possibility that the parser uses highly detailed information during processing because this would be too costly in computational terms.

The results of the unrestricted corpus set revealed that PPs attached to the preceding VP (61.4%) more frequently than to the preceding NP. This frequency pattern seems to correspond to native Greek speakers’ attachment preferences for temporary ambiguous PP structures; the results of the online self-paced reading task showed that Greek native speakers read VP attachments faster than NP attachments on the critical and post-PP segment. Thus, the results of the present study provide evidence that frequency patterns match parsing decisions on a coarse-grained level and appear to verify the predictions of the Tuning Hypothesis (Mitchell et al. 1995).
The extent to which frequency patterns guide parsing decisions, however, is to be further explored.

4.2.1 The role of the preposition

The analysis of the written and spoken corpus samples showed that attachment frequency varied across prepositions me, se, apo and ja; although me-, se- and apo-PPs attached more frequently to the preceding VP, ja-PPs attached significantly more frequently to the preceding NP. In addition, se-PPs differed significantly from me- and apo-PPs because of their stronger VP attachment. These results seem to match the on-line task results mainly in terms of me- and apo-PPs. Both in the corpora and in the psycholinguistic task, me- and apo-PPs were found to have a similar pattern of attachment; VP attachments were more frequent and were read faster than NP attachments.

On the other hand, a discrepancy between the corpora and the psycholinguistic data was found in sentences with se- and ja-PPs. The VP attachment preference that was found in the corpus data for sentences with se-PPs did not match the readers’ on-line reading time data which did not exhibit preferences for either attachment site. The reason for this difference between corpus data and participants’ online attachment preferences seems to be the highly underspecified nature of preposition se. Being almost ‘devoid of content’ (see e.g. Anagnostopoulou, 2005; Horrocks & Stavrou, 2007), preposition se very much depends on the meaning of other constituents in the structure, and especially the verb; this is most likely to be the reason why se-PPs were found to have such a high frequency of VP attachment in the corpora. On the other hand, it is very possible that when participants were presented with segmented sentences in the online self-paced reading task, and started building a structure and creating expectations about the incoming material, se did not give them enough information so as to make a definite parsing decision. If these assumptions hold, it appears that the results of the present study do not support the claims of constraint-based models (e.g. McRae, Spivey-Knowlton & Tanenhaus, 1988), connectionist models (e.g. Rohde, 2002) or probabilistic models (e.g. Jurafsky, 1996) which argue – to a greater or lesser extent – that parsing decisions are solely guided by frequency. It rather seems to be the case that parsing decisions are made ‘on the spot’ on the basis of structural but also lexical semantic information about the main constituents of the structure.

4.2.2 The role of definiteness

The experimental material of the online self-paced reading task included sentences in which the NP complement of the verb was always definite and the NP prepositional object was either definite or indefinite. This manipulation was conducted in order to examine the possible effect of Definiteness Agreement in Greek native speakers’ online parsing decisions. Some evidence for Definiteness Agreement was provided in the ILSP corpus analysis; even though the percentages of definite-definite VP and NP attachment sentences were both very high, the percentages of definite-indefinite NP attachment were lower than those of VP attachment in sentences with me- and se-PPs. In addition, the corpus data showed that definite-indefinite NP attachments were less frequent in sentences with me- and se-PPs than in sentences with apo- and ja-PPs. These results corresponded to participants’ online reading times; definite-indefinite NP attachments with me- and se-PPs were read significantly slower than definite-definite NP attachments. This means that participants were garden-pathed upon encountering me- and se-PPs (but not apo- and ja-PPs) with indefinite NP objects. In
this sense, it could be argued that the comprehension data were not essentially different from the production data. It should be noted, however, that the fact that Definiteness Agreement is a grammatical requirement makes it more possible to be found both in production and comprehension data.

5. Conclusion
Parsing preferences were found to generally correspond to corpus frequencies on a coarse-grained level; Greek native speakers’ faster reading times for VP attachment biased sentences did actually correspond to the VP attachment advantage in the corpus data. These data can be argued to be consistent with the Tuning Hypothesis (Mitchell et al. 1995) which posits that corpus frequencies are expected to be reflected on online parsing preferences. Nevertheless, the results of the present study showed that at the level at which the different prepositions were taken into account, corpus frequencies did not fully correspond to native Greek speakers’ online preferences. Even though participants’ online preferences for me- and apo-PP sentences patterned with corpus frequencies, there was no such correspondence for sentences with se- and ja-PPs. The examination of an even finer level of analysis at which animacy was taken into account revealed that the online VP attachment preference for ja-PPs matched with the relevant frequency data. On the other hand, participants’ online preference for neither attachment site in sentences with se-PPs, was not found in any of the grain sizes that were examined in the corpus data. Thus, even though the present study showed that there is correspondence between corpus frequencies and online comprehension preferences in most of the grain sizes that were taken into account, it cannot be claimed that it is actually corpus frequencies that guide parsing. If we accept that frequency is all there is in parsing, then the online patterns of attachment would be left unexplained. Greek native speakers’ online parsing data provide support for the claim that parsing decisions are made ‘on the spot’ on the basis of cues provided by the input. Under this line of reasoning, it is claimed that lexical cues such as the choice of preposition in the PP site play a definite role in participants’ attachment decisions.

References


