

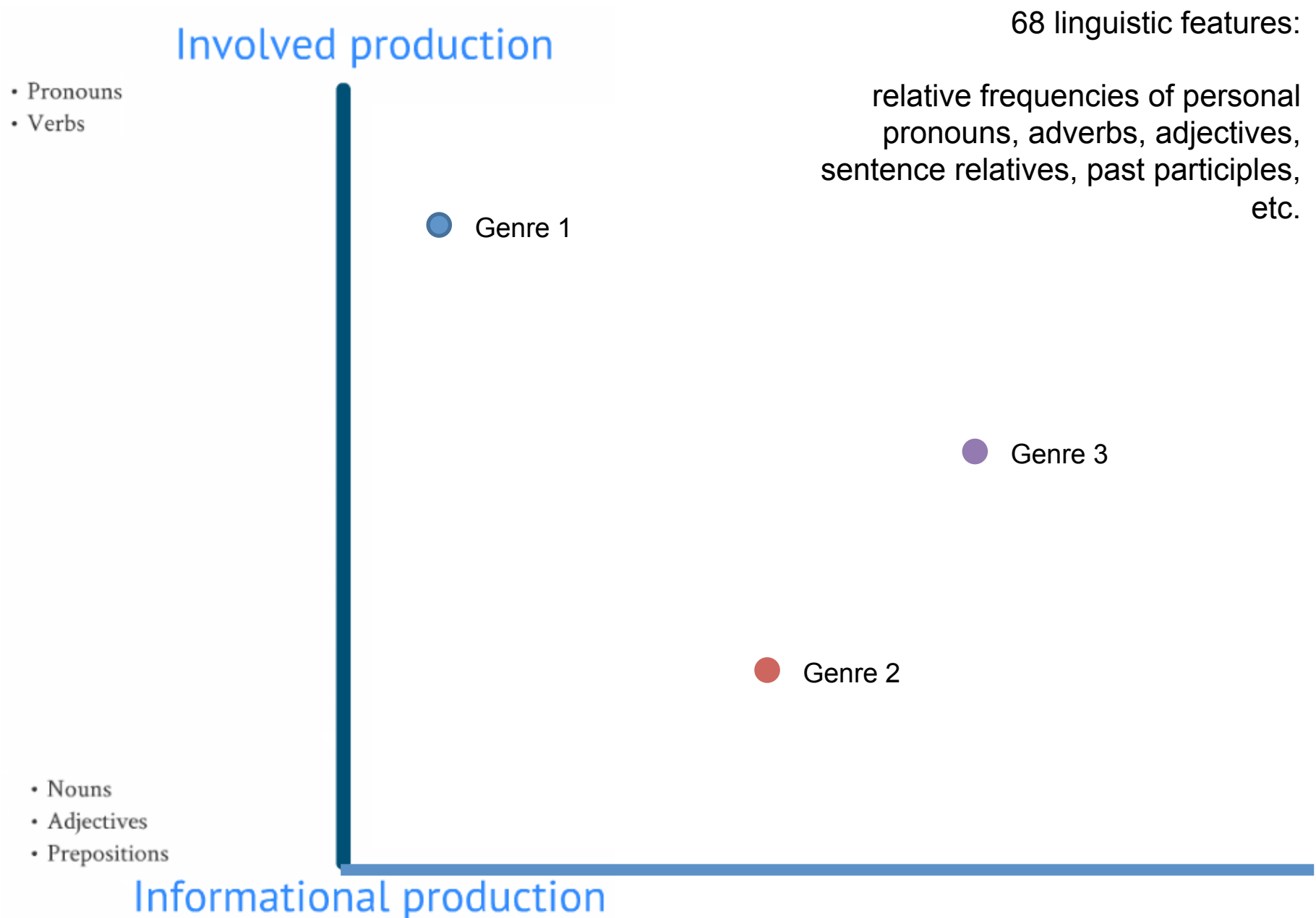
The *Multidimensional Analysis Tagger*, or how I stopped worrying and created a tagger using regular expressions

Dr Andrea Nini

andrea.nini@manchester.ac.uk

2nd BAAL Corpus Linguistics SIG event 2016

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.



Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Dimension 2

Narrative

Fiction

Other genres

Non Narrative

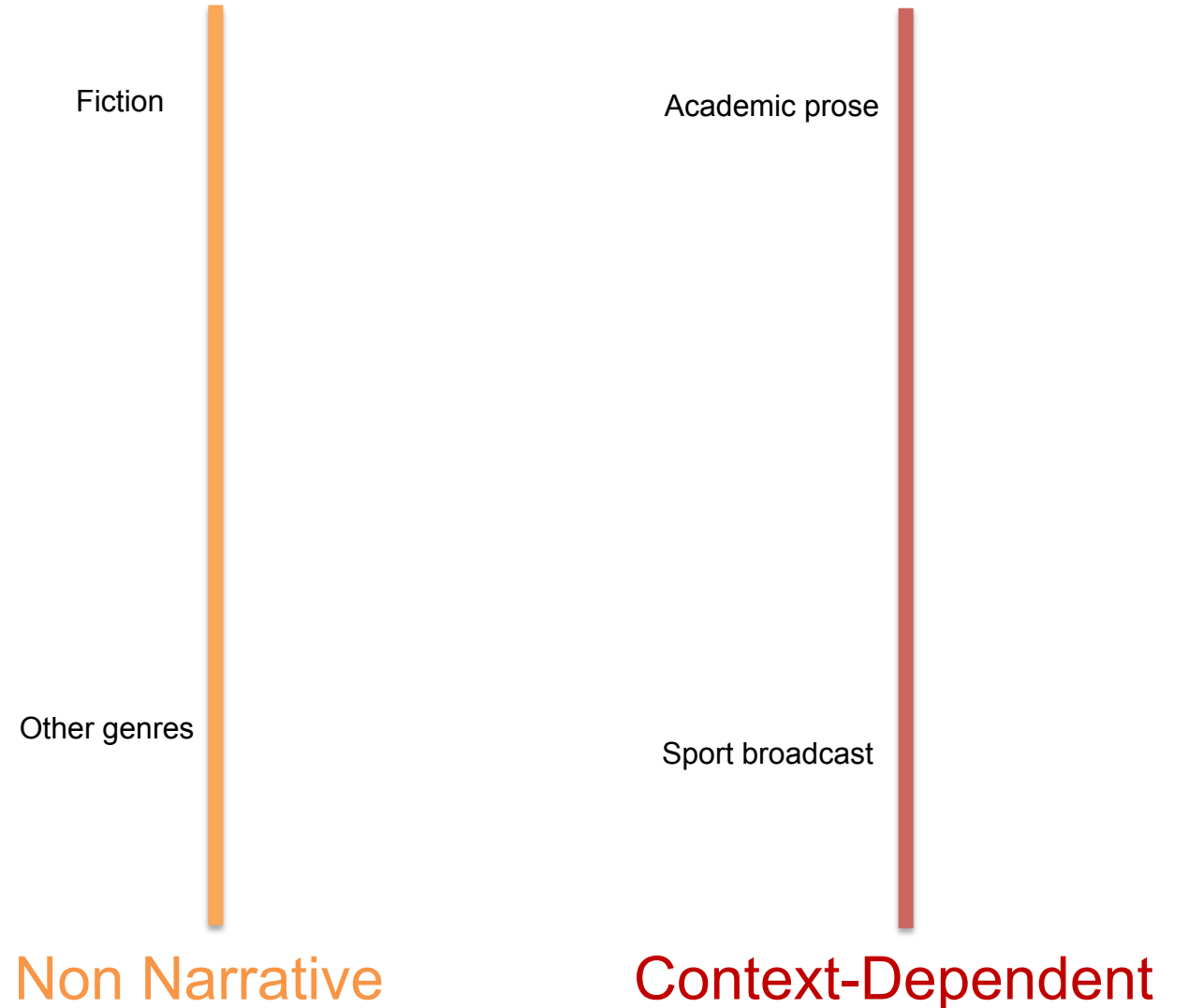
Dimension 3

Context-Independent

Academic prose

Sport broadcast

Context-Dependent



Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Dimension 4

Overt Expression of Persuasion

Professional letters
Speeches

Other genres

Dimension 5

Abstract

Academic prose

Conversation


Non-Abstract

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

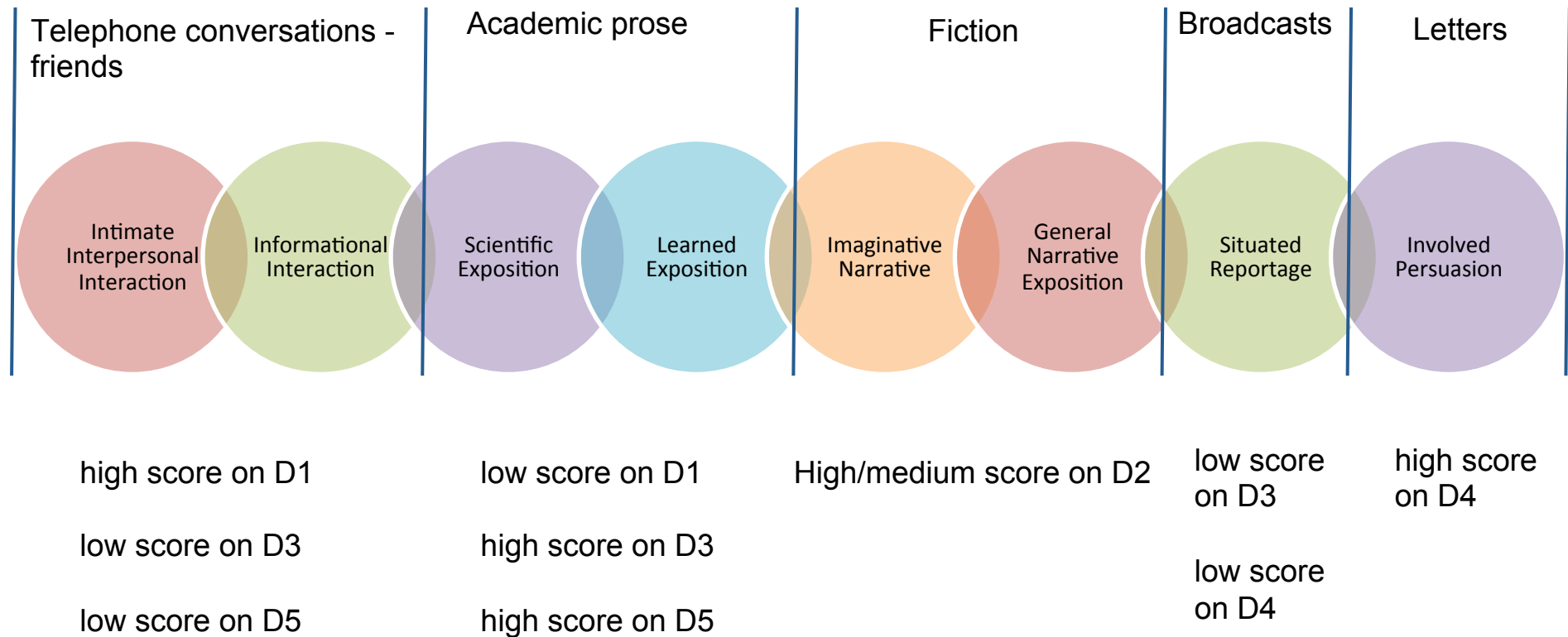
Dimension 6

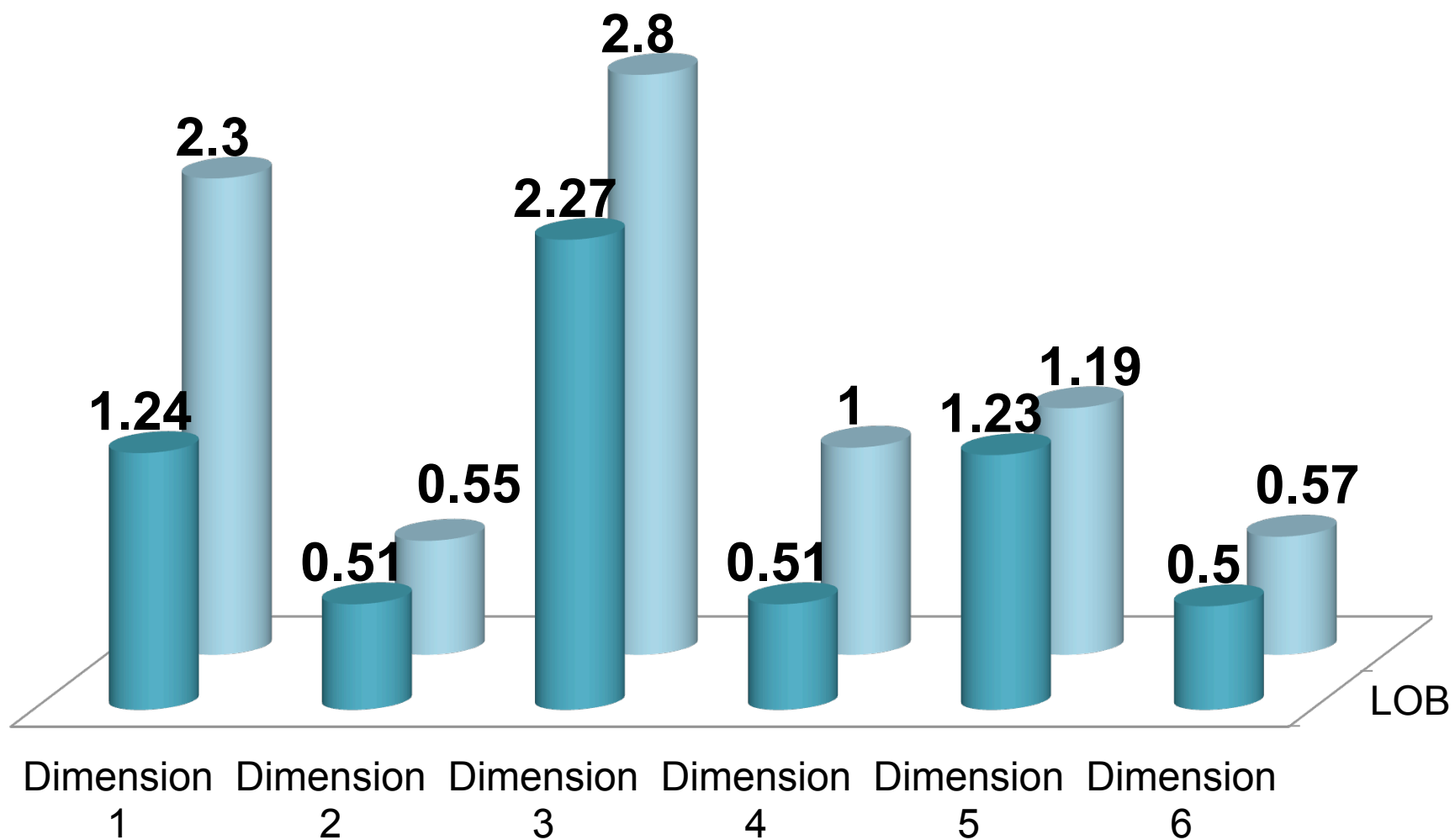
On-Line Informational Elaboration

Speeches

A thick vertical red line is positioned to the right of the word 'Speeches', extending from the top of the word down towards the bottom of the slide.

Biber, D. (1989). A typology of English texts. *Linguistics*, 27(1), 3–43.





16. **total other nouns**

All nouns included in the dictionary, excluding those forms counted as nominalizations or gerunds.

This count provides an overall nominal assessment of a text. Nominalizations and gerunds are excluded from the total noun count so that the three features will be statistically independent. In addition to Wells (1960), overall noun counts have been used by Carroll (1960) and Blankenship (1974).

(F) **PASSIVES** (nos. 17–18)

Passives have been taken as one of the most important surface markers of the decontextualized or detached style that stereotypically characterizes writing. In passive constructions, the agent is demoted or dropped altogether, resulting in a static, more abstract presentation of information. Passives are also used for thematic purposes (Thompson 1982; Finegan 1982; Weiner and Labov 1983; Janda 1985). From this perspective, agentless passives are used when the agent does not have a salient role in the discourse; *by*-passives are used when the patient is more closely related to the discourse theme than the patient. Studies that have used passives for register comparisons include Carroll (1960), Blankenship (1962), Poole (1973), Poole and Field (1976), O'Donnell (1974), Marckworth and Baker (1974), Ochs (1979), Brown and Yule (1983), Young (1985), Chafe (1982, 1985), Chafe and Danielewicz (1986), Biber (1986), and Gail (1986).

17. **agentless passives** 18. ***by*-passives****

(a) BE + (ADV) + (ADV) + VBN + (*by*)**

(b) BE + N/PRO + VBN + (*by*)** (question form)

(** no. 18 with the *by*-phrase)

(G) **STATIVE FORMS** (nos. 19–20)

Only a few studies have used stative forms for register comparisons. These forms might be considered as markers of the static, informational style common in writing, since they preclude the presence of an active verb. Conversely, they can be considered as non-complex constructions with a reduced informational load, and therefore might be expected to be more characteristic of spoken styles. Kroch and Hindle (1982) analyze existential *there* as being used to introduce a new entity while adding a minimum of other information. Janda (1985) notes that stative or predicative constructions (X *be* Y) are used frequently in note-taking, although the *be* itself is often dropped. Predicative constructions with *be*-ellipsis are also common in sports announcer talk (Ferguson 1983). These predicative constructions might be characterized as fragmented, because they are typically

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

The enemy was defeated by our troops.

The_DT enemy_NN was_VBD defeated_VBN by_IN
our_PRP\$ troops_NNS

(is|are|am|was|were|being|been)_\w+ \w+_VBN by_

[BYPA]
tagging

By passive clauses in Text 1: +1
counting

Chafe and Danielewicz (1980), Biber (1986a), and Grabe (1986).

17. **agentless passives** 18. **by-passives****

(a) BE + (ADV) + (ADV) + VBN + (*by*)**

(b) BE + N/PRO + VBN + (*by*)** (question form)

(** no. 18 with the *by*-phrase)

```
#tags "by passives"
```

```
if ($word[$j] =~ /\b($be)/i) {
```

```
    if (($word[$j+1] =~ /_VBD|_VBN/ && $word[$j+2] =~ /\bby_/i) ||
```

```
        ($word[$j+1] =~ /_RB|_XX0/ && $word[$j+2] =~ /_VBD|_VBN/ && $word[$j+3] =~ /\bby_/i) ||
```

```
        ($word[$j+1] =~ /_RB|_XX0/ && $word[$j+2] =~ /_RB|_XX0/ && $word[$j+3] =~ /_VBD|_VBN/ &&  
        $word[$j+4] =~ /\bby_/i) ||
```

```
        ($word[$j+1] =~ /_NN|_NNP|_PRP/ && $word[$j+2] =~ /_VBD|_VBN/ && $word[$j+3] =~ /\bby_/i)  
        ||
```

```
        ($word[$j+1] =~ /_XX0/ && $word[$j+2] =~ /_NN|_NNP|_PRP/ && $word[$j+3] =~ /_VBD|_VBN/ &&  
        $word[$j+4] =~ /\bby_/i)) {
```

```
        $word[$j] =~ s/_(\w+)/_ $1 [BYPA]/;
```

```
    }
```

```
}
```

Corpus of Forensic Malicious Texts

about 40,000 words
about 130 texts

- 1. What are the most significant dimensions of variation and text types?**
- 2. How do these texts compare to other registers of the English language?**

Your employee has been kidnapped and will be released for a ransom of £175,000. With a little luck he should be still O.K. and unharmed, to prove this fact to you will in in the next day or so receive a recorded message from him. He will be released on Friday 31 January 1992, provided:

On Wednesday 29 January a ransom of £175,000 is paid, and no extension to this date will be granted.

The police are not informed in any way until he has been released.

On Wednesday 29th at 4pm (on line 0213582281) you will receive a short recorded message from the hostage. To prove he is still alive and O.K. he will repeat the first news item that was on the 10am, Radio 2 news. He will then give further instructions. A second and more detailed message will be given at 5.05 pm the same day. Your watch must be synchronized with the 5pm

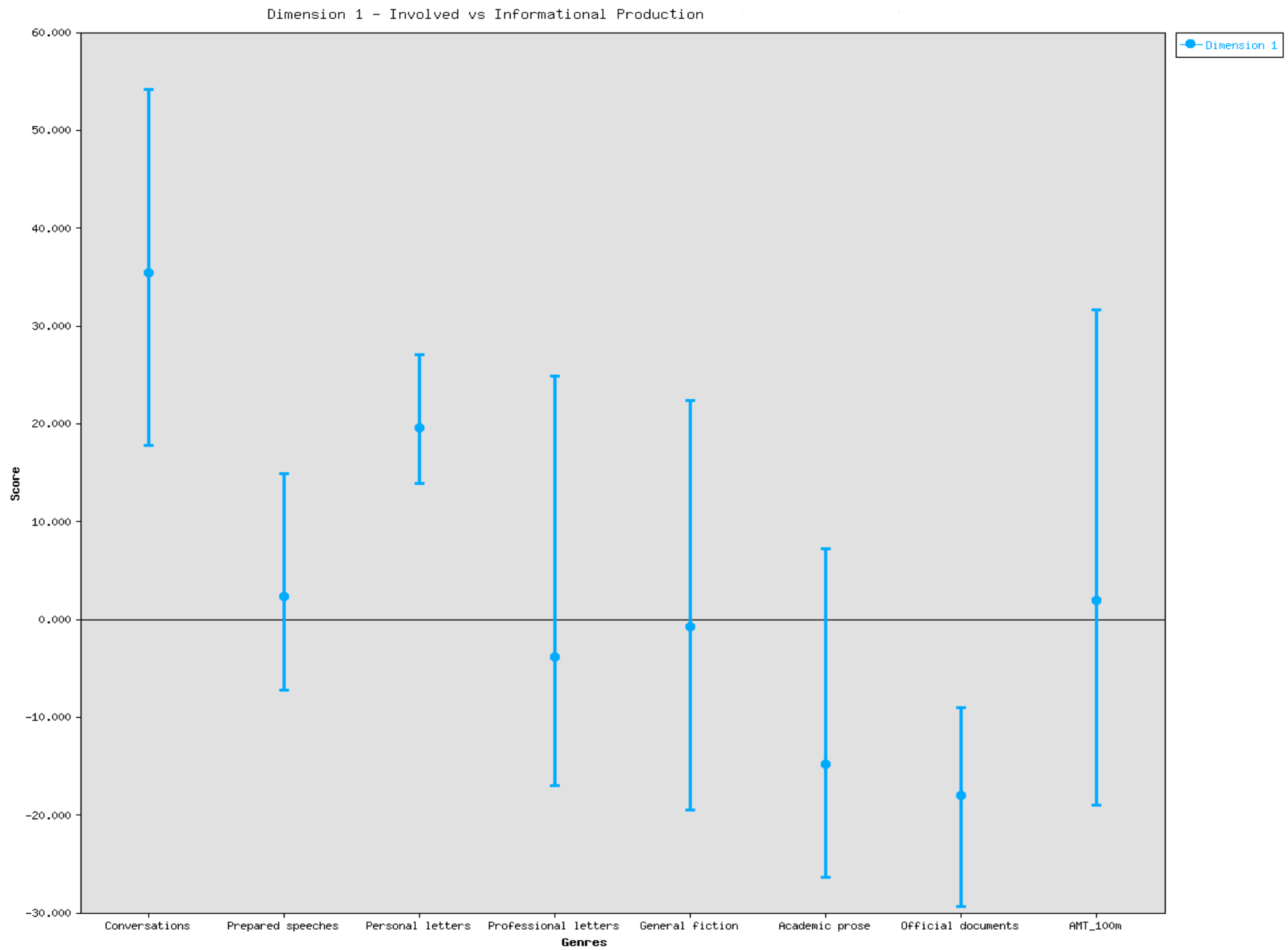
pips on Radio 2. The location of the second call will be given at 4pm, so transport with a radio must be available.

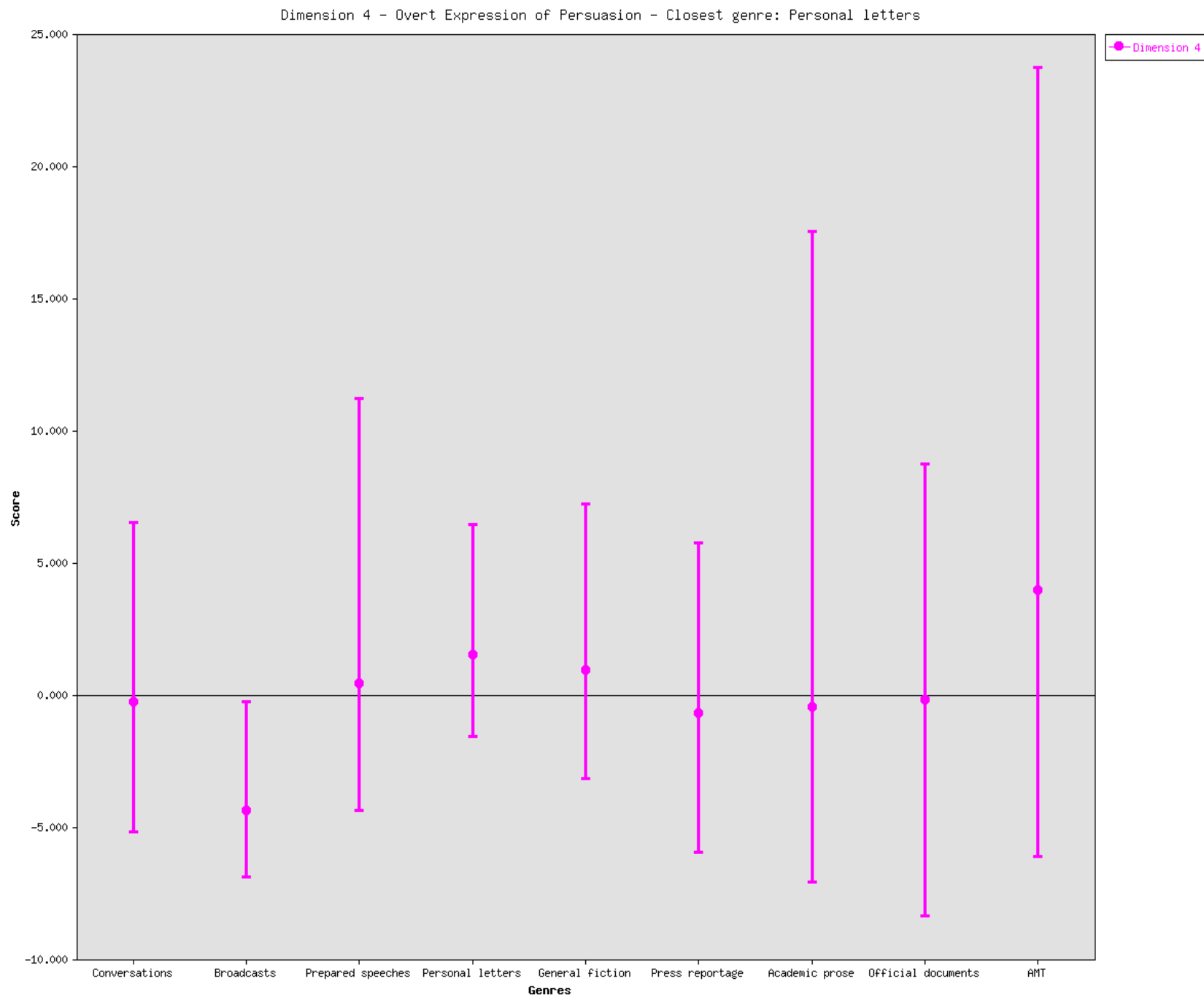
The money must be carried in a holdall and made up as follows, precisely; £75,000 in used £50. £75,000 in used £20. £25,000 in used £10 packed in 31 bundles, 250 notes in each.

Kevin Watts (if not the hostage) must be the person to receive all messages and carry the money to the appointed place.

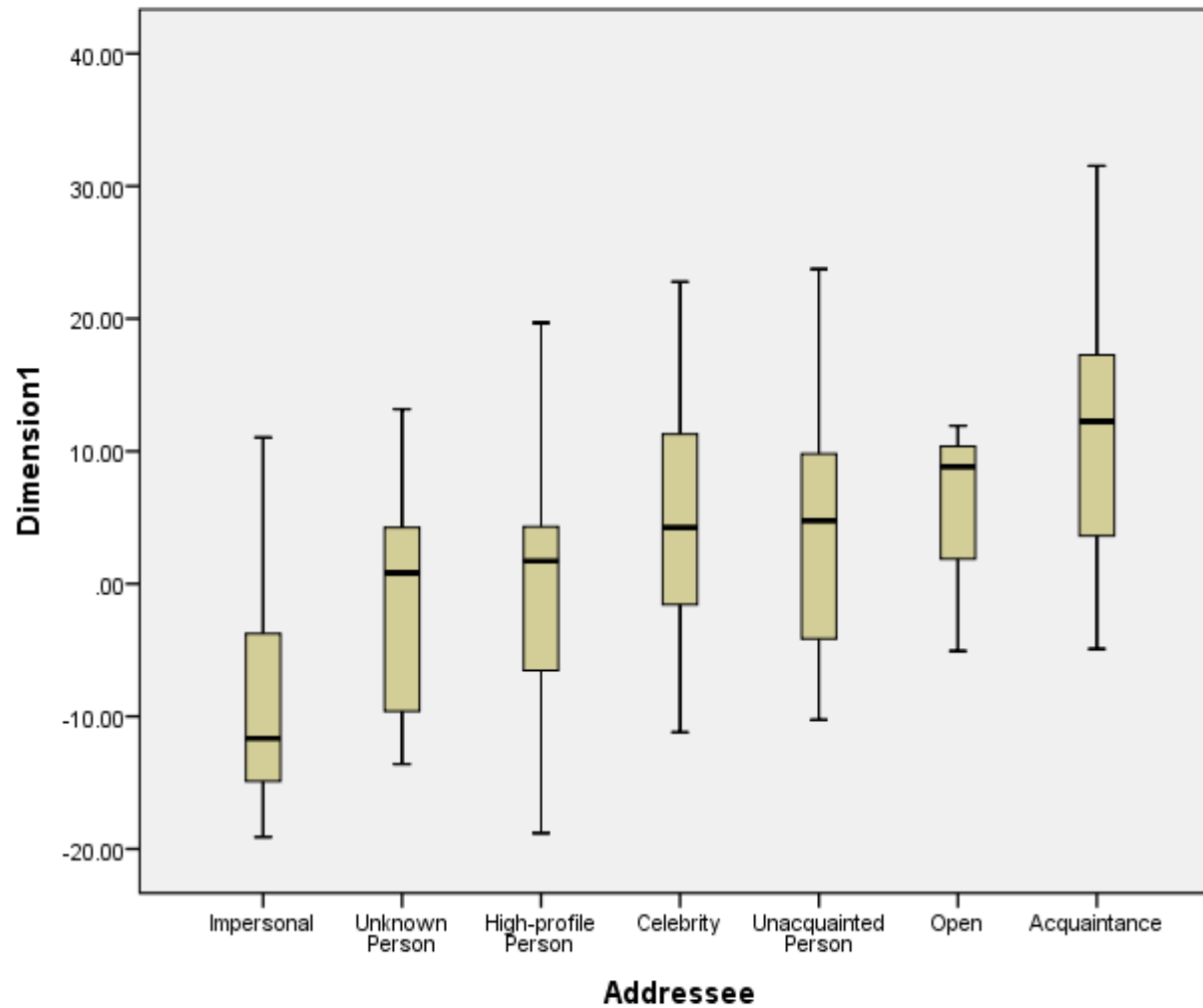
However, please note that all messages will be pre-recorded, so no communication or negotiations can be made.

YOU HAVE BEEN WARNED. HIS LIFE IS IN YOUR HANDS.

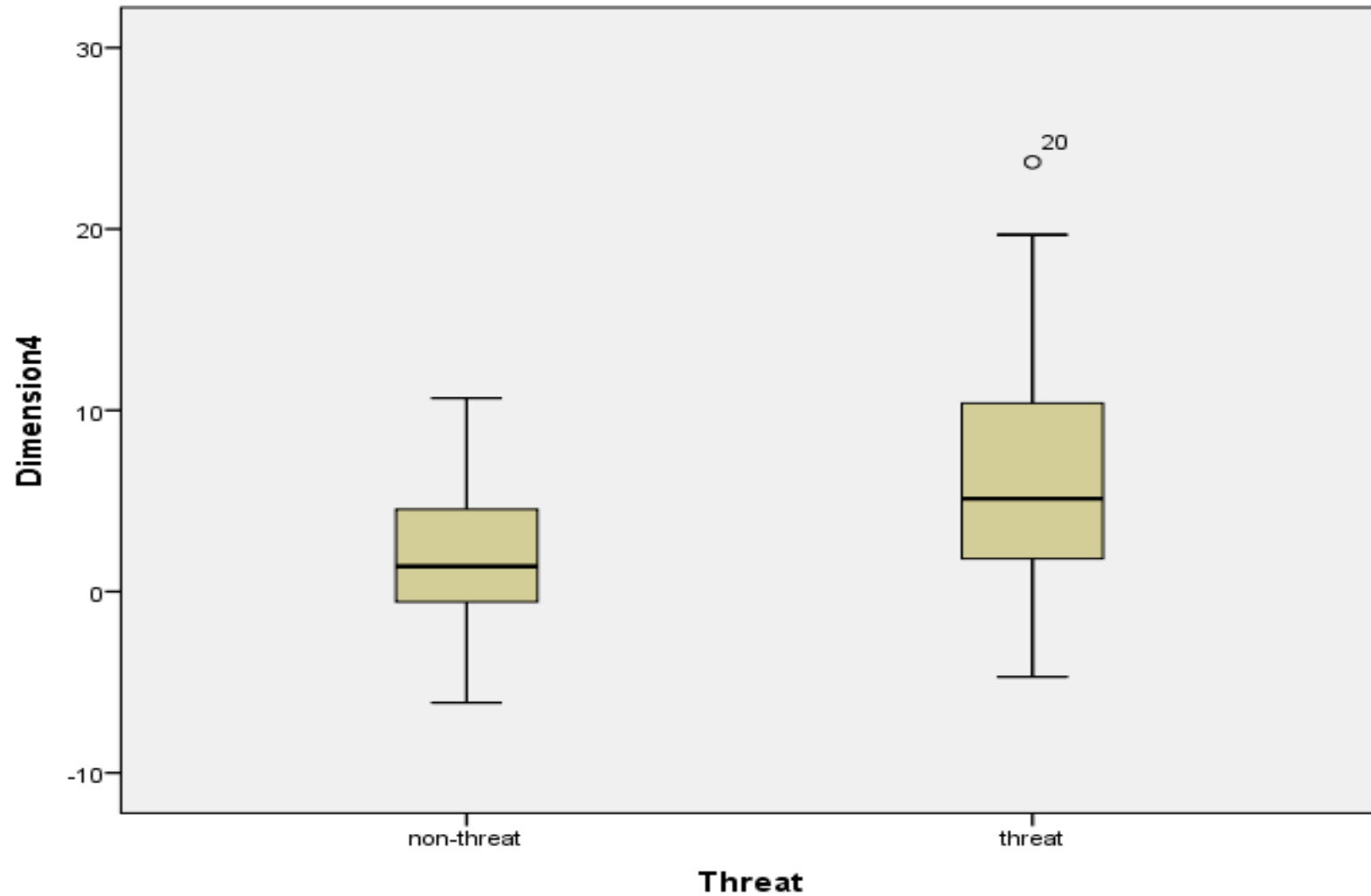




Addressee



Presence of threat



Conclusions



- ☐ **Biber's Dimensions and text types are very reliable and still useful**
- ☐ MAT can be used to exploit them
- ☐ Regexes can be combined with other powerful tools, such as taggers or parsers
- ☐ MAT can be used to analyse register variation when a factor analysis cannot be applied
- ☐ Stop worrying

The *Multidimensional Analysis Tagger*, or how I stopped worrying and created a tagger using regular expressions

Dr Andrea Nini

andrea.nini@manchester.ac.uk

2nd BAAL Corpus Linguistics SIG event 2016