Words that go together: An exploration of the idiom principle in institutional spoken English

Adriano Ferraresi (University of Bologna, Italy), Silvia Bernardini (University of Bologna, Italy) and Maja Miličević (University of Belgrade, Serbia)

1. Introduction

The notion of the idiom principle was introduced by Sinclair (1991:110) to account for collocations, or "semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments". Research in corpus linguistics and psycholinguistics has attempted to chart the extent of its applicability in different communicative settings and conditions, to understand the relationship between attestedness in corpora and formulaicity, or the property for a sequence of words to be "prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray 2002:9).

2. Aims

Taking advantage of a unique communicative setting (the European Parliament), our aim in this paper is to investigate the extent to which different groups of English speakers, engaging in spoken monologic discourse under different task constraints, appear to resort to the idiom principle. Our analysis focuses on lexical collocations and attempts to answer two questions: 1) are there differences in the number of collocations used by different groups of speakers under different task constraints? 2) is there evidence of faster/slower access, or more/less holistic processing, in the different conditions, as indicated by prosodic features? It has been suggested that "[f]ormulaic sequences tend to be uttered with particular prosodic features such as alignment with pauses and intonation units, resistance to internal dysfluency [and] no internal hesitations" (Wood 2015:23), and that such features can be taken as cues to the holistic storage of formulaic sequences (Dahlmann and Adolphs 2007). Here we focus on disfluency signals (silent and filled pauses and false starts) within collocations as evidence in favour or against operation of the idiom principle.

Our corpus is a subset of EPTIC (Bernardini et al. 2016), including 80 speeches delivered in February 2011, selected so as to include an equal number of read-out speeches and speeches delivered impromptu by native English speakers and speakers of ELF, English as a Lingua Franca.[1] Interpreted speeches (from Italian and French) are also included,[2] leading to the setup shown in Table 1.

[1] Even though MEPs have a right to speak in their native language, a few elect to speak English.

[2] Including interpretations from two different languages was meant to limit the possible influence of the source language, while ensuring that a larger range of interpreters was sampled. No attempt to measure the source language variable was made in this particular study.

	Native	ELF	Interpreted from FR	Interpreted from IT
Impromptu	10 (2,352)	10 (2,117)	10 (3,089)	10 (2,736)
Read-out	10 (3,398)	10 (5,366)	10 (2,959)	10 (2,475)
TOTAL	20 (5,750)	20 (7,483)	20 (6,048)	20 (5,211)

Table 1. Subset of EPTIC used, with n. of speeches (and n. of words) per subcorpus.

While the corpus is very small, it is highly comparable and apt to investigate authentic spoken English produced in a range of conditions. The comparison between impromptu and read-out speeches should highlight the different role played by the idiom principle in planned, written-to-be-spoken language vs. spontaneous, unscripted language (Erman and Warren 2000). The comparison between native and ELF speeches follows in the wake of extensive previous research on native vs. non-native (mainly learner) use of formulaic language (Kecskes 2007; Li and Schmitt 2010). Finally, simultaneous interpretations should allow us to observe if and how this condition of extreme cognitive effort, requiring a "roughly equal activation of two languages" (Sharwood Smith and Truscott 2014:208), impacts on the idiom principle.

3. Method

Collocation candidates are extracted based on corpus queries targeting the following structures:

- adjective+noun: e.g. *fair elections*
- noun+noun: e.g. *eyewitness accounts*
- verb+noun: e.g. *using violence*
- noun+verb: e.g. *industry faces*

We discard word pairs including proper nouns and numerals, and make sure that the remaining ones are syntactically well-formed (thus we retain the pair *use feed* in "*use* of contaminated *feed*", but discard *touch authorities* in "get in *touch* with the *authorities*").

To evaluate the collocation status of the remaining pairs, frequencies are obtained from ukWaC (Baroni et al. 2009) and used to calculate word association strength relying on two association measures (AMs), *t*-score (*t*) and Mutual Information (*MI*): these are known to emphasize different types of collocations, i.e. highly frequent vs. strongly associated ones (Durrant and Schmitt 2009). The cut-off point between collocations and non-collocations is based on the median of the bigram scores: $MI \geq 3$, and/or $t \geq 8$; bigrams with frequency <3 are excluded (cf. Evert 2008).

Disfluency signals within the selected collocations are identified as occurrences of silent and filled pauses, which are transcribed in EPTIC speeches as "…" and "ehm" respectively, as well as false starts (e.g. "sustainable m- management"). Prior to collocation extraction, the reliability of pause annotation was independently checked against the audio files by at least two authors.

Two kinds of statistical analyses are performed. In the first, the number of bigram tokens scoring high on a single AM (high-*t*, high-*MI*) and on both AMs (high-*MI*&*t*) is calculated for each speech and expressed as a percentage (e.g. of high-*t* combinations relative to the number of word combinations found in a speech). Percentages of each type of collocation are then used as an outcome variable in three sets of comparisons, which

are tested for significance using Wilcoxon rank sum tests: a) original vs. interpreted speeches, b) originals produced by native vs. ELF speakers, and c) originals delivered as read-out vs. impromptu speeches. In the second analysis, the presence vs. absence of a pause is used as a binary outcome variable in a logistic regression model with speech status (original/interpreted) and AM status (high-*t*/high-*MI*/high-*MI*&*t*) as categorical predictors. All analyses are carried out using *R*.[3]

4. Results

All comparisons returned non-significant differences, with one exception: read vs. impromptu speeches differ in terms of the percentage of high-*MI* combinations (W=117.5, p<.05). When the analysis is performed separately for the native and ELF groups, the difference is only significant for the natives (W=21.5, p<.05; Figure 1).

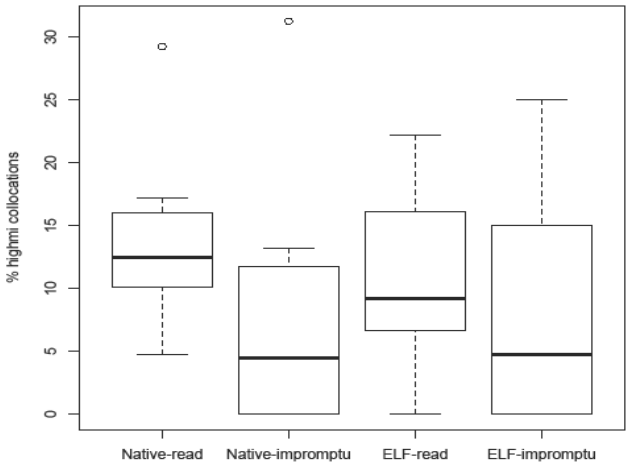

Figure 1. High-*MI* collocations in native and ELF speeches.

The regression analysis of disfluency signals (Figure 2) shows that, as might be expected, speeches interpreted into English contain more pauses overall than speeches originally produced in English. More interestingly, in both subsets of speeches the more robust collocations, those with high *MI* and high t-score, are less likely to contain a pause or a false start than those scoring high on a single measure. These predictors (AM and speech status) contribute significantly to the presence of disfluencies (coefficients in Table 2).

[3] http://www.r-project.org/

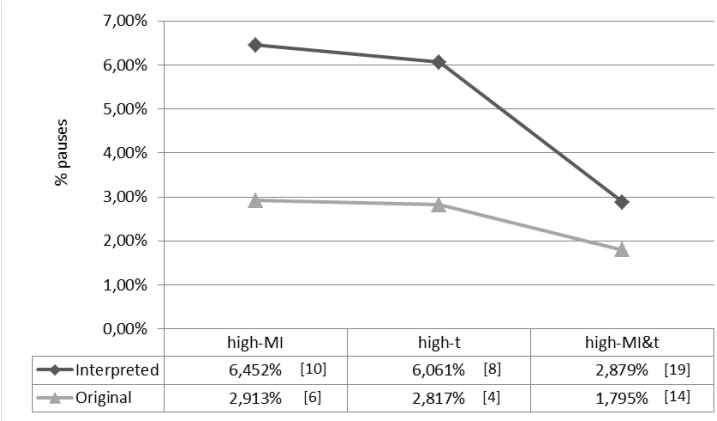

Figure 2. Percentages of pauses by speech and AM status, with frequencies in square brackets.

Fixed effect	**Coeff.**	**SE**	***Z***	***p***
(Intercept)	3.388	0.298	11.360	<.001
Speech status (interpreted)	-0.633	2.667	-2.377	<.05
AM status (high *t*)	0.046	0.391	0.117	ns
AM status (high *MI* & *t*)	0.702	0.311	2.257	<.05

Table 2. Summary of the logistic regression model.

5. Discussion and conclusion

This paper reports on a study focusing on the operation of the idiom principle in a small-scale yet closely comparable corpus of English speeches delivered at the EU Parliament by different sets of speakers and under different task conditions (read vs. impromptu, native vs. ELF, original vs. interpreted).

Our results concerning the number of collocations used suggest that ELF speakers in this international setting and specialized register do not differ significantly from native speakers in their use of both frequent and salient (high-*MI*) collocations, when improvising their speeches. These results would seem to lend partial support to Morgan's (2014:63) hypothesis that "MI sensitivity is not a marker of the so-called "native-speaker", but rather a high degree of proficiency with a particular register". However, the fact that native speakers use significantly more high-*MI* collocations in their prepared speeches than in those delivered impromptu, while no such difference is found in the ELF speeches, is coherent with the repeatedly observed higher sensitivity of native speakers to strength of association (Ellis and Simpson-Vlach 2009). At least as concerns our setting and the lexical collocations we have focused upon, the native speaker preference for strongly associated collocations seems to be more related to their "functional utility", than to the greater working memory demands associated with speech constructed in real time (ibid:62-3). This finding seems confirmed by the lack of significant differences between speakers and interpreters, despite the latter group's much heavier cognitive load (also signalled by the higher number of disfluencies in interpreted output overall). Our results concerning disfluencies within collocations lend support to the idiom principle hypothesis, since the more robust collocations, i.e. those that are both frequent and strongly associated, are delivered with fewer hesitations by all groups.

In further work we would like to enlarge the corpus and repeat the analyses for collocation types rather than tokens, to take into account the possibility that native

speakers use a wider variety of collocations than ELF speakers (Granger 1998). Finally, access to the data analyzed here will be provided through the NoSketch Engine platform. Speech-to-video alignment is being performed, giving access to the synchronized videos from concordance lines, an especially welcome feature for corpus studies of spoken language.

References

Baroni, M., S. Bernardini, A. Ferraresi & E. Zanchetta (2009). The Wacky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3), 209–226.

Bernardini, S., A. Ferraresi & M. Miličević (2016). From EPIC to EPTIC – Exploring simplification in interpreting and translation from an intermodal perspective. *Target*, 28(1), 61–86.

Dahlman, I. & S. Adolphs (2007). Pauses as an indicator of psycholinguistically valid multi-word expressions (MWEs)? *Proceedings of the workshop on a broader perspective on multiword expressions.* 49–56.

Durrant, P. & N. Schmitt (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, 47(2), 157–177.

Ellis, N. C. & R. Simpson-Vlach (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, 5(1), 61–78.

Erman, B. & B. Warren (2000). The idiom principle and the open choice principle. *Speech*, 20(1), 29–62.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics* (volume 2, pp. 1212–1248). Berlin and New York: Mouton de Gruyter.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 145–160). Oxford: Oxford University Press.

Kecskes, I. (2007). Formulaic language in English Lingua Franca. In I. Kecskes & L. R. Horn (Eds.), *Explorations in pragmatics: Linguistic, cognitive and intercultural aspects* (pp. 191–219). Berlin and New York: Mouton de Gruyter.

Li, J. & N. Schmitt (2010). The development of collocation use in academic speechs by advanced L2 learners: A multiple case-study approach. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 2–46). London: Continuum Press.

Morgan, J. A. (2014). *Explorations into the psycholinguistic validity of extended collocations.* Unpublished MA thesis. Portland State University.

Sharwood Smith M. and Truscott J. (2014). *The multilingual mind: A modular processing perspective*. Cambridge: Cambridge University Press.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Wood, D. (2015). *Fundamentals of formulaic language: An introduction.* London and Oxford: Bloomsbury Publishing.

Wray, A. (2002). *Formulaic Language and the lexicon*. Cambridge: Cambridge University Press.