ANGLICISMS IN ROMANIAN – DIACHRONIC CONSIDERATIONS

Arina GREAVU
Universitatea Lucian Blaga, Sibiu

Abstract

The paper investigates the quantitative impact of borrowed English words and phrases in a corpus consisting of eight years of the business and financial publication Capital. Several lines of analysis are pursued: the numerical evolution of simple versus phrasal Anglicisms, the distribution of borrowed versus codeswitched elements across frequency ranges, the internal complexity of phrasal Anglicisms. The perspective adopted is mainly a diachronic one, the article seeking to identify any existing trends in the outcomes of English-Romanian contact, although synchronic considerations are not entirely absent from the analysis. The results of this analysis are interpreted within the framework of some of the most influential theories put forth in the language contact literature.

Key-words: Anglicism, borrowing, codeswitch, word type, word token

1. Introduction

The perspective adopted in this study for the definition of the term Anglicism is a synchronic one. Thus, we consider the formal criterion to be of paramount importance in separating Anglicisms from Romanian words, this approach being in line with the definition given to the term by other researchers (Stoichițoiu-Ichim 2006, Görlach 2002, Onysko 2007), who regard as an Anglicism any word "recognized in its form (spelling, pronunciation, morphology) as coming from English" (Onysko, 2007: 90). Thus, an Anglicism can be defined as any English lexical element in the economic publication *Capital* that can be *formally* related to English. This definition excludes integrated loanwords which are English borrowings only in a historical sense, leaving only those recent English words and phrases which have not been yet adapted to the system of Romanian.

Following several influential contributions in the classification of different language contact phenomena, most notably in the field of the borrowing/ codeswitching dichotomy (Treffers-Daller 1994, Muysken 2000, Myers-Scotton 2002), length of constituency will be used as the main criterion for classifying Anglicisms in this paper. As a result, single words (including hyphenated compounds) and phrasal English importations in Romanian will be

analysed as distinct categories. The terms simple Anglicisms or borrowings will be used to describe the former category, while phrasal Anglicisms or codeswitches will be used in relation to the latter.

By including both single words and phrases under the conceptual coverage of the term Anglicism, we partly follow Avram (1997), who defines an Anglicism as "o unitate lingvistică (nu numai cuvânt, ci şi formant, expresie frazeologică, sens sau construcţie gramaticală) şi chiar tip de pronunţare sau/şi de scriere (inclusiv de punctuaţie) de origine engleză, indifferent de varietatea teritorială a englezei, nu doar din cea britanică." (1997: 11).

In the absence of rigorous methodological guidelines to separate Anglicisms from the bulk of the vocabulary, any attempt to do this will remain to a certain extent subject to personal interpretations, and therefore open to debate. The main purpose of the present study is not to make an exhaustive inventory of English-origin words in Romanian, but rather to conduct a quantitative and qualitative analysis of a certain subclass of these words, with a view to describing the ongoing contact between the two languages in question.

2. The corpus and data elicitation

The source of the corpus was the business magazine *Capital* on CD-Rom, consisting of Adobe PDF files. This raw data underwent a series of processing procedures, i.e. Optical Character Recognition, sentence splitting, tokenization and part-of-speech tagging and lemmatization[1]. The texts thus obtained, amounting to 20,262,068 tokens, allow for an efficient way of retrieving and processing Anglicisms. Dedicated software tools designed specifically for this project were used to tap the source of Capital 1998-2005.

The first stage of this process was the generation of decontextualized word lists showing all the individual word types in the corpus, and thus facilitating a faster identification of possible Anglicisms. According to the definition given to the term Anglicism in the previous section, a number of 4,495 word types and 63,175 word tokens were elicited from a corpus of 78,068 types (Capital 2005). The list resulting by subtracting the Anglicisms from the total was later used as a Stoplist blocking the occurrence of the component words from appearing in subsequent lists for the other years. The same data elicitation procedure was repeated for the seven years 1998-2004. The following table shows the results of this first stage of data elicitation.

[1] All these processing tasks were performed by Eckhard Bick (researcher) and Tino Didriksen (student assistant), from the Institute for Language and Communication (ISK) at the University of Southern Denmark. The tagging was done using the MSD tagger developed by the Research Institute for Artificial Intelligence of The Romanian Academy, under Professor Dan Tufiş' supervision. The pos-tagged corpus is available at http://corp.hum.sdu.dk/cqp.ro.html

Year	Types (Angl.)	Tokens (Angl.)	Total nr types	Total nr tokens
1998	2,220	30,738	73,610	2,035,220
1999	2,338	38,021	77,108	2,442,619
2000	2,310	42,449	71,925	2,342,260
2001	2,673	49,380	72,548	2,371,601
2002	2,917	54,941	77,104	2,653,352
2003	3,648	51,363	76,549	2,635,769
2004	3,755	55,739	78,066	2,889,367
2005	4,495	63,175	78,068	2,891,880
Total	8,148	385,806	209,647	20,262,068

Table 1. Number of Anglicisms in Capital 1998-2005 (unfiltered frequencies)

Deciding whether a given word was an Anglicism or not was problematic at this stage for two main reasons: firstly, the existence of homographs in English and Romanian (e.g. E. *deal* and R. *deal* 'hill', E. *sale* and R. *sale* as a pronoun, E. *fast* and R. *fast* 'pomp', to name just a few) and secondly, the use of English proper nouns, of original English works or quotations of English texts. In order to solve this ambiguity which could have resulted into prematurely and wrongly interpreting some words as Anglicisms context had to be taken into account. Thus, all the types identified initially had to be checked individually for their contextual usage, this filtering process having as a result the establishing of the actual token frequency for each Anglicism. For example, the unfiltered token frequency of *business* in 2005 is 413, while the word is used as an Anglicism, i.e. outside proper names, in only 321 instances. Various other words displayed a significant discrepancy between the filtered token frequency and the unfiltered one, *e.g. advertising, broker, consultancy, etc*.

This filtering process had as a result the establishing of the final amount of data, i.e. Anglicism types and tokens as shown in the table below. These figures include all individual English words appearing in the corpus, before compounds and phrases were identified and separated.

Year	Types	Tokens
1998	1,160	14,152
1999	1,439	20,082
2000	1,351	21,016
2001	1,656	26,553
2002	1,705	28,706
2003	1,819	23,364
2004	1,851	25,406
2005	2,135	27,928
Total	48,22	187,207

Table 2. Number of Anglicisms in Capital 1998-2005 (filtered frequencies)

We believe that these figures allow for diachronic conclusions concerning the recent numerical development of Anglicisms in *Capital*, providing some objective evidence to the commonly held belief that the number of Anglicisms in Romanian is increasing. Thus, the table above shows that the number of English words rose steadily from 1998 to 2005. However, these statistics include all individual lexical elements of phrasal constructions and of compounds listed as separate items, even though their occurrence might be restricted to certain phrases and compounds representing direct importations from English. For example, English function words are inherent elements of codeswitches, but they never appear as single Anglicisms outside this codeswitching environment. Similarly, many content words are restricted to phrasal Anglicisms. For example, *country* appears for 67 times in expressions such as *country manager* but is never found alone, while *head* has over 60 occurrences in expressions like *head hunting, head of corporate affairs, head of office* but not one individually. Due to the high number of such codeswitched units, the separate counting of these words would have considerably distorted the results of the final quantitative and qualitative analysis. This is why phrasal units had to be identified and counted separately. The results of this stage of the analysis are presented in Figure 1 below, which gives the evolution of borrowings (one-word Anglicisms) and codeswitches (two- and multi-word Anglicisms) over the studied period of time.

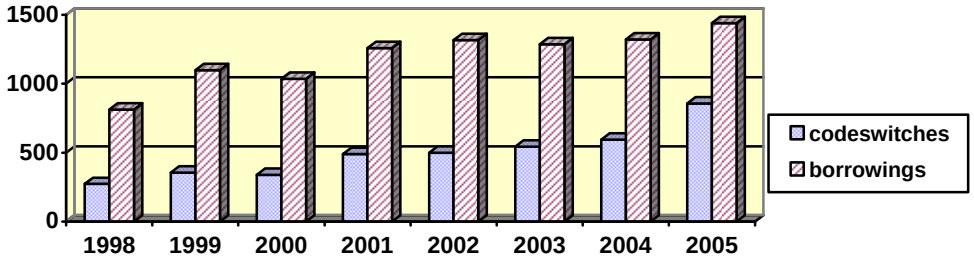

Figure 1. The evolution of borrowings and codeswitches in Capital 1998-2005 (types)

This chart shows an overall rise in the number of borrowed elements, but also a significant increase in the number of codeswitches from 1998 to 2005. Moreover, the two types of elements did not have a proportional growth, as phrasal Anglicisms seem to have had a more noticeable increase than borrowings. More detailed quantitative aspects regarding Anglicisms will be discussed contrastively only for the years 1998 and 2005 in the following section.

3. Quantitative analysis of Anglicisms in the corpus

This section presents the quantitative results of Anglicisms for *Capital 1998* and *Capital 2005*. The discussion starts out with the total number of Anglicisms (types and tokens) and the overall distribution of their token frequencies in the two corpora. Thus, the total number of individual Anglicisms is 1,160 types occurring in 14,152 instances in 1998, while in 2005 there are 2,135 types and 27,928 tokens. Compared with the total number of types and tokens in the corpus, in 1998 the rate of Anglicisms amounts to 1.50 % of all the types and 0.58 % of all the tokens in the corpus, while in 2005 the proportion held by Anglicisms increases considerably to 2.76% of all types and 1.05% of all tokens. The quantitative impact English words had in *Capital 1998* and *Capital 2005* is presented in table 3 below.

	1998		2005	
	Types	Tokens	Types	Tokens
Total nr. of words	73,610	2,035,220	78,067	2,891,880
Total nr. of Anglicisms	1,160	14,152	2,135	27,928
Percentage of Anglicisms/ total nr. of words	1.57	0.69	2.73	0.96

Table 3. Number of individual Anglicisms in Capital 1998 and Capital 2005

These results show an average of one Anglicism for every 63.45 word types and one Anglicism for every 172.41 word tokens in 1998. This average increases quite significantly in 2005, when one Anglicism is used for every 36.56 word types and for every 103.54 tokens. The table above also reveals a difference between the numerical proportion Anglicisms hold in the total in terms of types, and their representation as far as tokens are concerned, in both years under consideration. Such a discrepancy indicates a low repetition rate for these words, which is due to their new, socially unadapted character in Romanian, but also to the fact that most of the borrowed elements are content words.

In a study on Anglicisms in German, Onysko (2007: 114) finds a situation similar with the one resulting from the present research, and explains it as being a consequence of the types of elements borrowed. Thus, since content words have a lower frequency of occurrence than function words, and since the latter type of words are not borrowed at all in his corpus, being restricted exclusively to instances of codeswitching, the average frequency of

occurrence of borrowed words is relatively low. We believe we can use the same argument to show why only about every 173rd word in the corpus of 1998 and every 104th word in 2005 is an Anglicism in token terms.

A more accurate account of the quantitative impact of Anglicisms in the two years of the studied corpus can be obtained by detailing the figures in table 3 so as to include the separate classes of single and phrasal Anglicisms. At this stage of the research we hypothesize that these two categories behave differently from a morphosyntactic perspective, and they should therefore be studied separately. This is why, in the rest of this paper the analysis will proceed separately for each of these two classes of Anglicisms. In detail, table 4 below presents the distribution of simple and phrasal Anglicisms in 1998 and 2005.

	Simple Anglicisms (borrowings)		Phrasal Anglicisms (codeswitches)	
	Types	Tokens	Types	Tokens
1998	812	11,863	273	648
2005	1,442	20,534	860	2,497

Table 4. Anglicisms in Capital 1998 and Capital 2005 by structural type

The statistics above show a significant rise in the number of codeswitches from 1998 to 2005, both as regards number of their individual types and the token frequencies these types had in the corpus. However, this increase should be interpreted with caution, mainly because the 2005 corpus is significantly larger than the 1998 one, i.e. by about 40%. In order to eliminate the statistical distortions resulting from this disparity and give the true dimensions of this evolution, we have calculated the representation of these two separate classes as percentages of a total. The results of this stage of the analysis are presented in the pie-charts below.

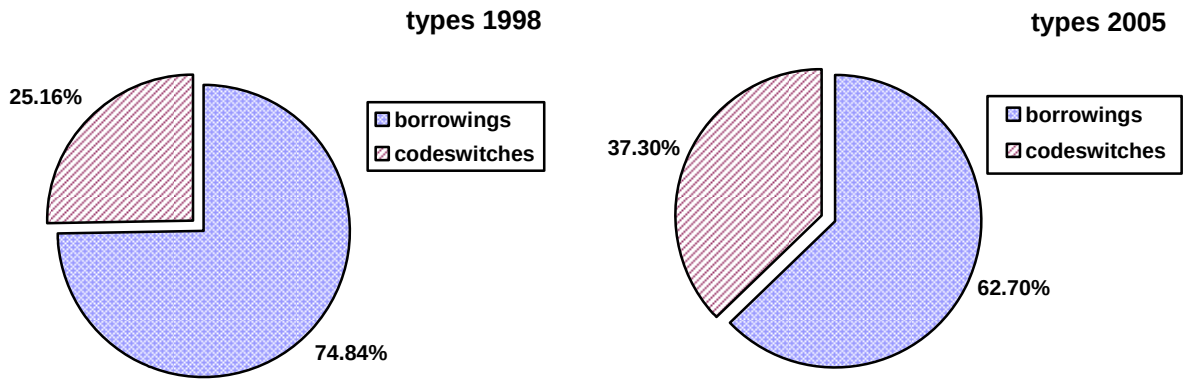

Figure 2. Distribution of Anglicisms according to form in Capital 1998 and Capital 2005 (types)

The increase in codeswitched elements from 25.16% of the total in 1998 to 37.30% in 2005, means that the use of phrasal Anglicisms rose by more than half in eight years. Thus, if in 1998 on average every 4th borrowed element was a two- or multi-word phrase, in 2005 every 3rd or even 2nd borrowed element was longer than one word. We believe this situation can allow us to speak of a change being underway as regards the pattern of language mixing in Romanian-English contact. The specialized literature generally recognizes a correlation between the level of proficiency in the source language and the occurrence of longer, more complex elements being transferred from this language. Myers-Scotton (2002) summarizes this correlation as follows:

When the overall prevailing pattern includes many bilingual CPs (with many mixed constituents), singly occurring forms (typically nouns) prevail. If speakers employ relatively many Embedded Language islands, they seem to be among the more proficient speakers. That is, it seems that higher language proficiency in the Embedded Language is necessary to feel at home producing islands. (Myers-Scotton 2002: 148)

CPs are formalized expressions of constituents in bilingual speech, while EL islands broadly coincide with the acceptation codeswitches have been given in this study. Myers-Scotton illustrates the proposal of proficiency-length of constituency correlation with quantitative evidence regarding the mixing patterns of two groups in an urban township of multilingual Black South Africans. Thus, the more educated and presumably more English proficient speakers of this community produced twice as many Embedded Language islands as compared to the less educated group, who used a much higher number of single foreign lexemes in their bilingual discourse.

However, it should be said that multi-word codeswitched elements of this type are not the expression of maximal proficiency in the source language, but represent an intermediary stage between single word insertions and intersentential switching, or switching between sentences. This idea was put forth and tested by Backus (1996), in a study of different generations of Turkish immigrants to the Netherlands, who adhered to one or the other of the three patterns of mixing, i.e. single words, EL islands and EL sentences, according to the level of bilingualism they had reached. We believe that the rather marked increase in Romanian – English codeswitching as evident from Figure 2 above testifies to an increasing level of English proficiency among the writers of the magazine in particular and possibly of other groups the Romanian speakers as a whole. This, in turn, could trigger further changes in the language mixing patterns involving the two languages in question. Although the studied

corpus cannot be expected to provide instances of true inter-sentential codeswitching, this being limited to the insertion in discourse of highly formulaic expressions of the type *big is beautiful, no comment, don't do this at home, expect the unexpected, it's a man's world, size is everything, trend is your friend,* it would be interesting to study the incidence and characteristics of this type of codeswitching in spoken language.

In order to further investigate this assumed transition from simpler to more complex Romanian-English mixing patterns, we have tried to analyse codeswitches in terms of their internal complexity. The contrastive analysis of the number of words serving to form such phrases in Capital 1998 and Capital 2005 has shown that in both cases more than half of all codeswitched elements were two-words phrases, while the remainder were made up of three or more words. However, the 1998-2005 period saw a 6% increase in the number phrasal Anglicisms using three or more elements. We believe that this increase can be correlated with the numerical growth of codeswitches as such, being the result of the same set of factors discussed above, most notably increased bilingualism among the writers of the magazine. Figure 3 below shows the results of this analysis.

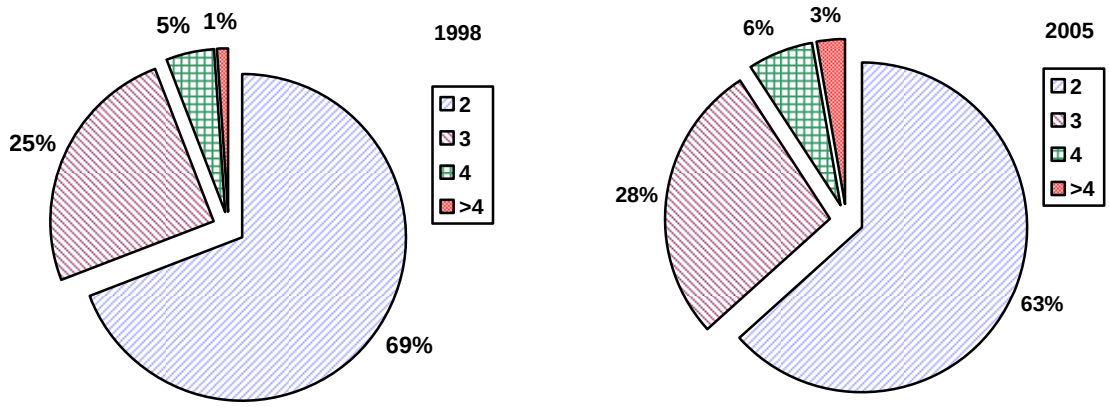

Figure 3. Classification of codeswitches according to length in Capital 1998 and Capital 2005

Another quantitative aspect of the contrastive study of Anglicisms in 1998 and 2005 was their frequency of occurrence. The distribution of codeswitched elements in terms of their frequency of occurrence as well as the evolution of these distributional patterns from 1998 to 2005 has been standardized for one million words in order to allow for comparisons between the two corpora. Figure 4 below shows no significant changes in the way phrasal Anglicisms were distributed across different frequency ranges in the two years studied. Thus, even if the number of phrases which appear only once in the corpus went down from 85% in 1998 to

78% in 2005, the number of phrases having a frequency of two or three occurrences increased from 10% to 16%, so that the broader category of codeswitches having a frequency occurrence of one-three has remained remarkably stable. The same stability is shown by codeswitches used for more than 4 times, their number remaining low at 4% in both years under consideration.

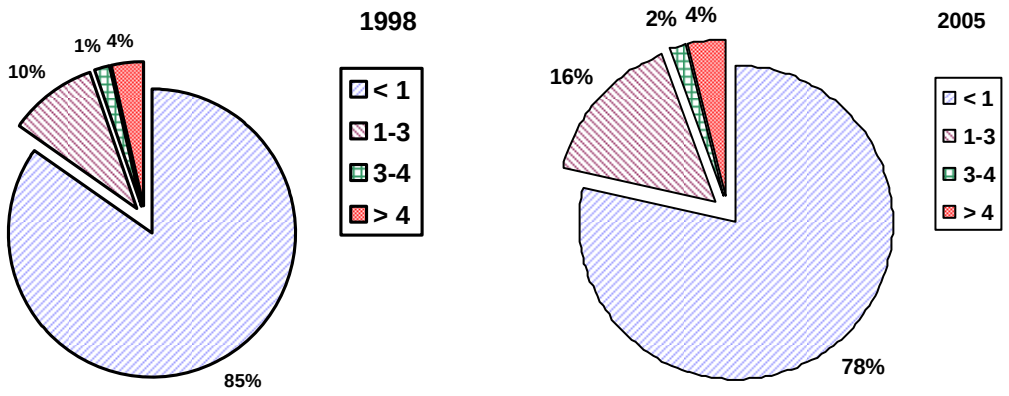

Figure 4. Distribution of codeswitches according to token frequency in Capital 1998 and Capital 2005 (rounded values)

The general conclusion seems to be that an overwhelmingly large proportion of codeswitched elements in Capital are used accidentally, while a very limited number of them have gained currency and become somehow established in the language. For example, out of a total of 860 phrasal Anglicisms in Capital 2005 only six are used more than 50 times (*art director, managing director, middle management, prime time, internet banking* and *media planner*), and only 45 more than 10 times. Such high incidence of speech borrowings usually testifies to growing intensity of contact manifested as increasing levels of bilingualism among recipient language speakers as well as favourable attitudes towards the source language, seen as a source of symbolic power, prestige or fashion. If we assume an increasing level of English proficiency among Romanian speakers, then we could expect singly occurring English words and phrases to become increasingly prominent in quantitative terms. The fact that no significant changes in this direction happened over an eight year period might suggest the idea that some aspects of language contact phenomena are more sensitive than others to changes in the social conditions surrounding the contact. More exactly, increased levels of bilingualism seem to be most strongly correlated with elements such as complexity and length of borrowed elements than with lexical diversity of transferred words and phrases.

Turning now to borrowings, these elements account for more than 70% of all English material in 1998, and for slightly more than 60% of all Anglicisms in 2005. When we

calculated the frequency of occurrence for these elements, we preferred to use lemmas rather than types because we believed this method would allow a better comparison with the class of phrasal Anglicisms. For example, the lemma/ type ratio for this class of Anglicisms in 2005 is only 1.09, which indicates a very low inflection rate for multi-word English elements. In other words, the vast majority of types are represented by the actual lemmas, rather than by inflected forms. Simple Anglicisms, on the other hand present a different situation: in 2005 there are approximately 900 lemmas and more than 1,400 types.

The results of this lemma-based distribution of simple Anglicisms across frequency ranges is presented in figure 5 below, standardized for one million words.

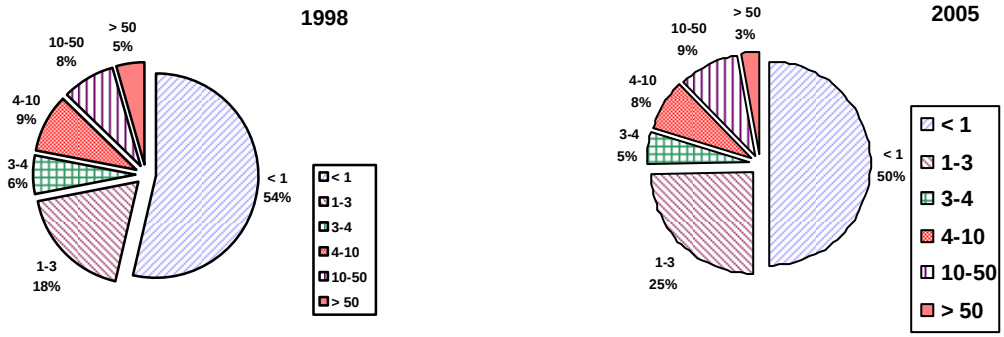

Figure 5. Distribution of simple Anglicisms according to token frequency in Capital 1998 and Capital 2005 (rounded values)

From a static, synchronic point of view, it is evident from the charts above that borrowings, just like codeswitches, show a very high "mortality rate" translated as low frequency of occurrence in the studied corpus. Such a situation can be accounted for as being a consequence of the stability of the language as a morphosyntactic and phonologival system which tends to reject new entries that will create homonymy or confusion (Grosjean, 2001).

On the other hand, one can also notice some significant differences between the frequencies of codeswitches and those of borrowings, the occurrence of the latter tending to be significantly higher. This situation lends further support to those theories regarding integration which correlate multi-word foreign items with low diffusion and acceptance among recipient language speakers, while single borrowings are more frequently used and widely accepted (Poplack and Sankoff, 1984). Such a correlation seems to be supported by psycholinguistic factors like ease of learning and reproduction, which are highly dependent of length and structural complexity of the borrowed element.

4. Conclusions

To conclude our discussion on the quantitative impact of Anglicisms in the corpus of Capital 1998-2005, this paper has shown that present-day Romanian is faced with a very distinct upward trend in this phenomenon. Such a general increase in the number of borrowed English words is accompanied by a discernable shift from borrowing to codeswitching, or from simple words to phrasal importations. This shift could be indicative of some changing social conditions which form the backdrop of the Romanian/English contact.

Bibliography

1. Avram, Mioara. *Anglicismele în limba română actuală*. Conferință prezentată la Academia Română. Bucureşti: Editura Academiei Române, 1997
2. Backus, Ad. *Two in One: Bilingual Speech of Turkish Immigrants in the Netherlands.* Tilburg: Tilburg University Press, 1996
3. Görlach, Manfred (ed.). *English in Europe.* New York: Oxford University Press, 2002
4. Grosjean, François. *Life with Two Languages: An Introduction to Bilingualism*. Cambridge, Massachusetts, London: Harvard University Press, 2001
5. Muysken, Pieter. *Bilingual speech: A typology of code mixing.* New York: Cambridge University Press, 2000
6. Myers Scotton, Carol. *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. New York: Oxford University Press, 2002
7. Onysko, Alexander. *Anglicisms in German: Borrowing, Lexical Productivity, and Written codeswitching.* Berlin, New York: Walter de Gruyter, 2007
8. Poplack, Shana and David Sankoff. *Borrowing: the synchrony of integration.* Linguistics 22, 1984, 99-135
9. Stoichițoiu-Ichim, Adriana. *Aspecte ale influenței engleze în româna actuală*. Bucureşti: Editura Universității din Bucureşti, 2006
10. Treffers-Daller, Jeanine. *Mixing Two Languages. French-Dutch Contact in a Comparative Perspective.* Berlin, New York: Mouton de Gruyter, 1994
11. *Colecția CD Capital*. Bucureşti: Ringier România, 1998-2005

Arina Greavu is university lecturer with University “Lucian Blaga” of Sibiu. Her research interests are in the area of language contact study, contact induced language change and specialized languages.