EXPLORING DISTRIBUTIONAL SEMANTICS WITH THE SPACEXPLORER TOOL

ÁGOSTON TÓTH University of Debrecen

Abstract: Distributional studies quantify the similarity of words by collecting word co-occurrence frequency information from large text corpora. According to the distributional hypothesis, this similarity is a semantic phenomenon. This paper aims to introduce the basics of Distributional Semantics and a new tool, spaceXplorer, which facilitates distributional investigations by collecting co-occurrence information from a Wikipedia snapshot (with or without using linguistic annotation) and displays word similarity information through a convenient, interactive user interface.

Keywords: cognitive science, computational linguistics, distributional semantics, lexicon

1. Introduction

Distributional Semantics is a computational, practice-oriented, data-driven approach to representing meaning. Co-occurrence statistical information supplies empirical evidence about a word's general potential for replacing another word, which gives us the opportunity of measuring word similarity. According to the distributional hypothesis, this similarity is a semantic phenomenon.

Distributional Semantics is a very active research program in Cognitive Science. It is based on a structuralist view on meaning (with roots that can be traced back to Saussure and Harris, cf. Sahlgren 2008): Distributional Semantics focuses on what is internal to language and assumes that other aspects of meaning (e.g. reference) will also be reflected by language-internal phenomena or remain irrelevant for description. Approximating the meaning of words is carried out by assessing distributional properties as manifested in corpora.

A geometric procedure is commonly employed in Distributional Semantics to represent and compare meanings. Co-occurrence events between words are usually collected as numerical features in *feature vectors* that stand for words in a *vector space*. Meaning differences and similarities can then be conveniently represented and calculated in this vector space by working with the feature vectors. More details about this process will be provided in section 2.

As shown above, Distributional Semantics is bound with strong ties to Linguistics and Geometry. Computational linguists have also found the distributional methodology an efficient yet powerful way of acquiring semantic information about words. As far as language technology is concerned, some of the first vector-space applications included the task of finding relevant documents in Information Retrieval (Salton 1971). Question answering (e.g. Tellex et al. 2003) and document clustering (e.g. Manning et al. 2008) may be implemented in a similar way. Comparable systems have been developed for word sense disambiguation (Schütze 1998), thesaurus generation through automatized discovery and clustering of word senses (Crouch 1988, Pantel and Lin 2002) and

named-entity recognition (Vyas and Pantel 2009). Pennacchiotti et al. (2008) use Distributional Semantics in a cognitive semantic context: they propose a method for extending FrameNet's scope by covering more (potentially: frame-evoking) lexical items through distributional lexical unit induction.

Psycholinguistics also has a major role it Distributional Semantics as corpusderived and psycholinguistic data correlate (gained from human similarity judgements, cf. e.g. Miller and Charles 1991, and from semantic priming experiments, e.g. Pado and Lapata 2007).

Distributional Semantics is a powerful model that has been used in many scientific disciplines, but it has an empirical side that can only be researched with proper tools that can process large corpora and find co-occurrence events between words.

2. How to build a Vector Space Model (VSM)?

Systems designed to collect distributional information about words rely on a geometrical interpretation of the empirical data (Widdows 2004). Each target word is represented in a multi-dimensional space by a feature vector. Each position of the feature vector signals or counts the number of co-occurrences of the given target word with one of the context words we use for describing target items. For example, if the word *drink* is a target word, the word *tea* is among the context words and *tea* occurs 23 times in the close vicinity (in the "context window") of *drink*, then the vector element corresponding to the word *tea* (in the context vector describing the word *drink*) will be set to 23:

 $v_{drink} = \langle freq_1, freq_2, ..., 23, ..., freq_t \rangle$

where *v* is a feature vector that represents a target word in a *t*-dimensional space; *t* stands for the total number of context words.

Large corpora (20-50-100 million words or even more) are necessary for this type of investigation. "Raw", unprocessed corpora may be suitable for the task. In the presence of linguistic annotation, we can take additional details into consideration.

We can compare the distributions of the target words by carrying out calculations with their feature vectors. Here, I will limit the discussion to two basic methods for comparing vectors: we can measure vector distances (Figure 1) or the cosine of the angle between vectors (Figure 2). The latter promises the advantage of being able to avoid problems arising from vector length differences, which is useful, since length depends on the frequency of context words.





Figure 1. Vector similarity: distance

Figure 2. Vector similarity: cosine

3. The spaceXplorer program

Unlike 1st-order word co-occurrence phenomena (seen when words occur together, e.g. in idioms, compounds with open spelling, phrasal verbs, light verbs, collocations, etc.) which are possible to spot by collecting examples of words and listing them (a task readily supported by most concordancing programs), 2nd-order word co-occurrence phenomena, which are observed and quantified in distributional studies, are not directly visible to the naked eye. We need software tools that process text, find occurrences of target words and collect and statistically evaluate contextual information about each occurrence in context so that the degree to which *words occur with the same words* can be calculated.

<u>.</u>		spac	eXplorer	1.0 - Tóth Ág	goston 2014			×
ile Find								
SM setup _analysis C	oncordancer							
	Context words	8k	~	Target words	20000 mostfreq it	ter 🗸	Status / Compatibility list	
Stopwords	1: year_N / 861 2: new N / 806	46 79	^	year_N / 86146	5	^	school_N / 1.0 college N / 0.625	^
predefined (344)	→ 3: time_N / 693	198		time_N / 69398		university_N / 0.532		
Processing target	4: school_N / 6 5: city_N / 6544	5781 13		city_N / 65443	/01		community_N / 0.472	
Lemma + POS	 6: state_N / 616 7: united N / 616 	6: state_N / 61638			state_N / 61638 united_N / 61388		center_N / 0.461	
Corpus size to proces	s 8: use_V / 6029	7: united_N / 61388 8: use_V / 60292 9: world_N / 55011 10: include_V / 54281		use_V / 60292	use_V / 60292 world_N / 55011 include V / 54281		church_N / 0.451 academy_N / 0.449 city N / 0.447	
100 M	9: world_N / 55			world_N / 550 include V / 54				
VSM window size	11: make_V / 5	3404		make_V / 5340)4 53003		student_N / 0.441	
1+1	12: university_f	12: university_N / 52003 13: county_N / 48203		county_N / 48203		house_N / 0.438		
	14: area N / 46	698		area_N / 46698	3		hospital_N / 0.436	
Compute new VSN	1 16: team_N / 4	5595		team_N / 45595		education_N / 0.425		
	17: states_N / 4	15490 / 44142		states_N / 454	90 4143		schools_N / 0.422 road_N / 0.419	
19: history_N / 42018		history_N / 42018		hall_N / 0.418				
	20: family_N / -	41666 46		family_N / 416	566		include_V / 0.411 campus N / 0.405	
Weigting	22: retrieve_V /	40013		retrieve_V / 40	013		art_N / 0.403	
pPMI	 23: know_V / 3 24: game_N / 5 	9901 19565		know_V / 3990 game_N / 395)1 65		state_N / 0.398 club_N / 0.398	
Vector comparison	25: south_N / 3	18950		south_N / 389	50		county_N / 0.392	
05	✓ 26: km_N / 384	26: km_N / 38490		km_N / 38490		museum_N / 0.388		

3.1. Measuring word similarity using spaceXplorer

Figure 3

SpaceXplorer (Figure 3) is a text processing tool for distributional studies. Before we start processing the corpus, the following parameters must be set from a list of options available on the left side of the interface shown in Figure 3:

- stopwords: whether we want to exclude the most common (function) words from processing,

processing target: lemma or word form with or without POS labels,

- corpus size to process: up to 100 million words (using 10,000-word samples from the Wikipedia snapshot),

- VSM window size: the size of the moving window around the target word where the context words are detected.

Near the top of the interface, we get options for selecting the amount of *target words* and *context words* for processing. The corresponding lists of target and context words become visible immediately, together with corpus frequency data, as shown in the *Context words* and *Target words* fields of Figure 3.

The present form of the program is prepared to process the TC Wikipedia ("Tagged and Cleaned Wikipedia") corpus available at .">http://nlp.cs.nyu.edu/wikipedia-data/>.

Having set all these parameters, the user can start processing the corpus and collecting statistical information about the target words by pressing the "Compute new VSM" button. Processing time depends mainly on the amount of context words, target words and the corpus size. Running the program at its most time consuming settings requires about 24 hours on a single PC to complete. It can finish in much less time (sometimes in minutes) with more modest settings, which is useful for parameter tuning or when the program is run for demonstration purposes only. Crucially, the user can save the results and load them later; in this way, multiple investigations can be carried out without having to process the corpus again with the same parameters.

When corpus processing is over, we can select any target word from the *Target words* list: as a result, the *Compatibility list* field displays all (target) words in decreasing order of distributional similarity to the selected target. The degree of similarity is also displayed in the list. In Figure 3, for instance, the noun *school* is selected for analysis. The noun *school* appears at the top of the similarity list (compatibility score: 1.0), followed by the noun *college* (compatibility: 0.625) and the noun *university* (compatibility: 0.532). Compatibility scores are all relative to the distribution of the selected target word (*school_N*) and they show the level of freedom (0-1) with which a word can replace another word in the corpus using the list, a similarity measure and a weighting scheme must also be selected on the interface (in this case: *pPMI weighting* and *cosine similarity*; you can find these settings in the bottom left corner). These options can be freely changed at any time without having to recalculate the word-context matrix by parsing the corpus again.

3.2. Creating concordances

A concordancer locates occurrences of a search expression and lists them in context. Concordance listings can be used in general linguistic research, lexicography and language learning. Concordances were used before electronic computers were invented in the middle of the 20th century; the first concordances were compiled for the Bible. In many cases, producing a concordance took years or even decades, and the result was published in several volumes. The invention of electronic computers sped up the process of compiling a concordance considerably. The spread of home computers also made the process interactive: you do not need to print the concordance, but you can generate it on the spot. Moreover, you can sort the concordance or change the options so that the pattern you are seeking becomes readily observable.

With spaceXplorer, you can create a concordance listing for the words selected from the *Compatibility list* field, which is filled with data during a VSM experiment. The concordance function can only be accessed after a VSM exploration is complete.

🗉 spaceXplorer - Tóth Ágoston 2014 – 🗖 🔀						×			
<u>File</u> <u>Select</u>									
VSM setup/analysis Concordancer									
			1						
Co	ontext words	20000 🗸	Target words	8000 most freq	~	Status / Compatibility list			
Stopwords	: new / 114032 : first / 90449	^	place / 26428 club / 26109	2	^	iowa / 0.086 community / 0.085		^	
predefined (144) V 3	: one / 8/832 - de / 67217		river / 25991	2		cave / 0.085			
Processing target 5	two / 62162		both / 25972		- 1	montana / 0.085			
Word form M	: united / 6209)	became / 2568	30		cross / 0.085			
7:	: time / 60149		country / 2559	5		towards / 0.085			
Corpus size to process 8	: city / 59456		before / 25519	before / 25519		smith / 0.085			
50 M Y 1	: school / 5809)	including / 25	including (25/15		names / 0.000			
1	1: national / 55	243	best / 25318			minnesota / 0.084			
VSM window size	2: world / 5461	0	government /	25237		riverside / 0.084			
1+1 v 1	3: states / 5190	9	album / 25132			grass / 0.084			
14	4: university / 4	19658	february / 251	27		tom / 0.084			
1	5: external / 49	254	day / 24895			view / 0.084			
Compute new VSM 1	6: links / 48522		station / 24/24 air / 24142			interstate / 0.084			
1	7: years / 47583	}				tributary / 0.084			
1	8: more / 4708	2	system / 24041			grand / 0.084			
1	19: american / 46822		now / 24038			henry / 0.084		_	
2	20: john / 46397		TOOTDAIL / 24017			bank / 0.084			
Weigting	1: county / 461	45	early / 23928			Iron / 0.083			
vergening 2	2: south / 45/0	4	al / 23914			band / 0.083			
pPMI V 2	23: see / 45/00		II / 23839			windsor / 0.083			
2	24: up / 44/90		main / 23/15			broad / 0.083			
Vector comparison 2	o: references /	44570	website / 2304	9		madicon / 0.002			
cos y 2	b: north / 4354	0	public / 25002			chorps / 0.000			
2	7: auring / 424	28	unusn / 23483	NCC		cheny/0.000			
2	8: history / 414	57	home (22210	324		London (0.003			
2	9: name / 413/	U	information /	כגרכר		dovor / 0.083			
3	u: over / 41341		ardar (22120	23243		Lover / 0.003			
		*	order / 23120		*	Id / ViVoo		~	

Figure 4

Figure 4 shows a compatibility list generated for the target word *river* by processing 50 million words (using no lemmatization and no POS labels at this time). After the analysis, the target words (the 8000 most frequent words of the corpus) are listed in descending order of compatibility with *river* in the *Compatibility list* field. Note that you do not see the beginning of the list in Figure 4, as the word *bank* has already been selected for further processing in this case. Also note that you do not have to browse the list manually: the program can help you locate target words and compatibles by using the corresponding functions on the *Select* menu.

When you activate the concordancer function by clicking on the *Concordancer* tab on the screen, you will get a *keyword in context* concordance shown in Figure 5. The major output elements are tabulated into columns:

- *left context*,
- *keyword* (bank),
- right context,
- degree of relevance (0.0 1.0),
- context words available in the given sentence for calculating relevance.

(p) [dec) (P) decimal (Construct) Under some (Construct) (c) the construct (Construct) (c)	We provide model with the first of the fi	application of the second seco	
Determined (Determined (Determ	Managements (construction) (construction) is the provide provide provide provide provide and experiment of the provide provide and experiment of the provide provide provide and experiment of the provide provide provide provide and experiment of the provide provi	Ele Select	
<pre>Dest Not n</pre>	Line to make 1 (1) For any intervent of a second of the stand arrow of the stand arrow of a second of the stand arrow of the stand ar	VSM setup/analysis Concordancer	
<pre>mining in the provide and decknows is in the part of the static parts of provide parts of the static parts of the static</pre>	<pre>aid has created using and adaptivity in the state approved person of person of person of the state approved person of person of the state approved person of</pre>	Lines to collect: 100 v Ctx words match: 8 v Make concordance Relative to: "river"	
<pre>sel has created roads and checkpoints in the Vest Band vich the stated purpose of preventing the uninhibited movement of :1.0; 1 [Report To IDF FM detrained to implement Vest Band and inch the stated purpose of preventing the uninhibited movement of :1.0; 1 Laboratory receives CAP certification for blood band] [September Awarded Grand Frize Best Entrepreneur Frize in</pre>	<pre>sel has created roads and checkpoints in the Vest Bank vith the stated purpose of preventing the uninhibited movement of 10 Laboratory receives CAP certification for blood bank Laboratory receives CAP certification for blood bank anaged to reach the Americans at the Elbe s west bank anaged to reach the Americans at the Elbe s west bank from behind his own met he used the boards to bank from behind his own met he used the boards to bank on ark and since its establishment as a bank the bank of a river sighted an energy machine you can be bank of a river sighted an energy machine you can be bank of a river sighted an energy machine you can be houring the position of director of the Reserve Bank of a river sighted an energy machine you can be houring the position of director of the Reserve Bank of a river sighted an energy machine you can be houring the position of director of the Reserve Bank of a river sighted an energy machine you can be houring the position of director of the Reserve Bank of a river sighted an energy machine you can be houring the position of director of the Reserve Bank of a river sighted an energy machine you can be houring the position of director of the Reserve Bank of a river sighted an energy machine you can be and the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the</pre>	set he created rooks and checkpoints in the Best has with the stards purpose of prevention the unlabilities accesses of 1. Advances the stards of the stards of the stard of the stard of the stard of the stards of the stard of the s	1.0. iiradi crested book seet istated purpose preventing koreant 200 preventing the set of the set o
1] Reports IDF PM determined to implement Vest Bank pullout says Israel not interested in toppling Hanas Haaretz] :1.0 / 95 hass of the Stonehenge monument a circular earth bank and dith dates to C E] [The Chide into the East Stonehenge monument a circular earth bank and dith dates to C E] [The Chide into the East Stonehenge monument a circular earth bank and dith dates to C E] [The Chide into the East Stonehenge monument a circular earth bank and dith dates to C E] [The Chide into the East Stonehenge monument a circular earth bank and the puck which rolled into the East Stonehenge monument a circular earth bank in the puck which rolled into the East Stonehenge monument is a bank the bank is base assets has surpassed billion] [The bank is 0.9 / 10	c)] Reports IDF PM determined to inplement Vest Bank pullout says Israel not interested in toppling Hanas Haarstz] 10 9 shase of the Stonehenge monument a circular earth bank and ditch dates to c BC] [The CAide 10 9 shase of the Stonehenge monument a circular earth bank and ditch dates to c BC] [The CAide 10 9 shase of the Stonehenge monument a circular earth bank and ditch dates to c BC] [The CAide 10 9 for the the bar share the Elbe vest bank but many more did not among the the Swelish Company 10 9 for the bank of a river sighted an energy machine guin on the 10 9 m aark and since its establishent as a bank the bank of a river sighted an energy machine guin on the 10 9 m aark and since its establishent as a bank the bank of a vest bank of a river sighted an energy machine guin on the 10 9 m aark and since its establishent as a bank the bank of a set rest sighted an energy machine guin on the 10 9 m aark and since its establishent as a bank the bank of the vest bank	sel has created roads and checkpoints in the West Bank with the stated pu	urpose of preventing the uninhibited movement of (1.0;
L Laboratory receives CAP certification for block bank [] [September Awarded Grand Prize Best Entrepreneur Prize in [0.95] hasse of the Stonehenge monument a circular earth bank and ditch dates to c BC] [The CAide	L Laboratory receives CAP certification for blood bank] [September Awarded Grand Prize Best Entrepreneur Prize in 9 9 shase of the Stonebenge monuent a circular earth bank and ditch dates to c BC] [The CAIde 9 9 for abelind his own net he used the boards to bank the puck which rolled into the Bast swacted net 9 9 for abelind his own net he used the boards to bank the puck which rolled into the Bast swacted net 9 9 for abelind his own net he used the boards to bank the puck which rolled into the Bast swacted net 9 9 for abelind his own net he used the boards to bank the puck which rolled into the Bast swacted net 9 9 nark and since its establishment as a bank the bank of a river sighted an energy machine guin on the 9 9 nack and since its establishment as a bank the bank is a base assets has suppassed billion] [The bank 0 9 ncluding the position of director of the Reserve Bank of New Zealand] [In he was ade a 1 9 9 ncluding the position of director in the state state the state state the state state the state state state the state sta	2] [Reports IDF PM determined to implement West Bank pullout savs Israe	el not interested in toppling Hamas Haaretz] ;1.0;
base of the Stomehenge monument a circular earth bank and dith dates to C B] [The Ckide	phase of the Stonelenge monument a circular earth bank and ditch dates to CBC] [The CAide 0 9 9 from behind his own net the Elles west bank but any more did not anong the the Swelish Company 0 9 from behind his own net he used the boards to bank the puck which rolled into the East s vacated net 0 9 more and and since its establishment as a bank the bank is base assets has surpassed billion] [The bank 0 9 more did not anong the the steril of the start s vacated net 0 9 more and and since its establishment as a bank the bank is base assets has surpassed billion] [The bank 0 9 more did not anong the the steril of the start start start and start is the start of the start start start and start start start start start start and start start start start and start star	Laboratory receives CAP certification for blood hank 1 [September Avai	rded Grand Prize Best Entrepreneur Prize in :0.99
<pre>sanaged to reach the Americans at the Elbe s west bank but many accredition to among them the Svedish Company [0 97] from behind his own met he used the boards to bank the puck which rolled into the East s vacated net [0 96] Etaing France a patrol reconnoiting on the west bank of of a river sighted an energy machine gymen on the [0 97] an ark and since its establishment as a bank the bank is base assets has surpassed billion] [The bank [0 97] including the position of director of the Reserve Bank (of Rev Zealant) [In we adde a [0 97] and the position of director of the Reserve Bank (of Rev Zealant) [In we adde a [0 97] and the position of director of the Reserve Bank (of Rev Zealant) [In we adde a [0 97] and the position of director of the Reserve Bank (of Rev Zealant) [In we adde a [0 97] and the position of director of the Reserve Bank (of Rev Zealant) [In we adde a [0 97] and the position of director of the Reserve Bank (of Rev Zealant) [In the state of the intervent for the state of the st</pre>	<pre>sensed to reach the Americans at the Elbe's west bank but many more did not among them the Swedich Company 0 9 from behind his course the west the board of a river sighted an ensay machine guin on the 0 9 from behind his course the stabilishment as a bank the bank of a river sighted an ensay machine guin on the 0 9 from between the stabilishment as a bank the bank of a river sighted an ensay machine guin on the 0 9 from between the stabilishment as a bank the bank of a river sighted an ensay machine guin on the 0 9 from between the stabilishment as a bank the bank of a river sighted an ensay machine guin on the 0 9 from between the stabilishment as a bank the bank of a river sighted an ensay machine guin on the 0 9 from between the stabilishment as a bank of the Reserve Bank of New Zealand] [In he was made a</pre>	phase of the Stonehenge monument a circular earth hand and ditch dates to	n c BC 1 [The Cǎide 0.98
from behind his own net he used the boards to bank the puck which rolled into the East s vacated net. 0996 Etaing France s patrol recompositing on the vest bank of a river sighted an energy machine guing on the 9096 Including the position of director of the Reserve Bank of New Zealand] [In keys a made a	From behind his own net he used the boards to bank the puck which rolled into the East s wacated net (0.9) Etaing France a patron reconnoiting on the west bank of a river sighted an energy machine guin on the 0.9 In mark and since its establishment as a bank the bank is base assets has surpassed billion] [The bank 0.9 In mark and since its establishment as a bank the bank of key Zealand] [In he was add a 0.9 We should be position of director of the Reserve Bank of Key Zealand] [In he was add a 0.9 Of the position of director of the reserve the bank of the set of the	anaged to reach the Americans at the File swest hand but many more did	not among them the Swedish Company :0.97
Etaing France a patrol reconnoitring on the vest bank of a river sighted an energy machine gun on the 0 94 nark and since its establishment as a bank the bank of a river sighted an energy machine gun on the 0 94 nark and since its establishment as a bank the bank of a river sighted an energy machine gun on the 0 94 nark and since its establishment as a bank the bank of the vest base assets has surpassed billion] [The bank 0 94 nark and since its establishment as a bank the bank of the vest base assets has surpassed billion] [The bank 0 94 nark and since its establishment as a bank the bank of the vest base assets has surpassed billion] [The bank 0 94 nark and since its establishment as a bank the bank of the vest base assets has surpassed billion] [The vest base as a set base assets has surpassed billion] [The vest base as a set base asset base as a set base as	Etaing France a patrol reconnoitring on the west bank of a river sighted an energy machine gun on the 10 9 purchased as ince its establishment as a bank the bank is base assets has surpassed billion 1 (The bank 10 9 purchased	from behind his own net he used the boards to hank the nuck which rol	lled into the Fast s vacated net :0.96
n mark and since its establishent as a bank the bank is base assets has surpassed billion] [The bank 0 94 including the position of director of the Reserve Bank of New Zealand] [The kase adda a 10 92 The start of the star	n mark and since its stabilishent is a bank the bank is base assets has surpassed billion 1 [The bank 0 9 n mark and since its stabilishent is a bank the bank is base assets has surpassed billion 1 [The bank 0 9 including the position of the ord intervent of the bank of the stabilishest of the stabili	Ftaing France a patrol reconnoitring on the yest hand of a river sighter	d an enemy machine gun on the
<pre>pin fund that the project is easy to set of increases of the set of the</pre>	<pre>unincluding the position of director of the Receive Band of New Zeelsand [] [In he wandes a to be add a to</pre>	in mark and cince its actablichant as a bank the bank s back accets has	surpassed billion 1.[The bank
And your set an	<pre>https://www.setup.com/setup.com</pre>	including the position of director of the Reserve Rank of New Zealand 1	In he was made a
and the control was defined by the last in	An an analysis of whether the function of the second field in the field in the second	We the posterior of the bold of the starte back of new zearand j	
BOW once contriduce by annual to be anticipated by anticipate by anticip	I Son after building Logichtin at the bank of Christian Creat Burlage we late the bank of Christian area the bank of Christian Creat Burlage we late the bank of Christian Creat Burlage we late the bank of Christian Creat Burlage and the bank of Christian Chreat Burlage and the bank of Christian Creat Burlage and the bank of Christian Chreat Burlage and the bank of Chreat	<pre>ref Bound Theory and the Massa of Labora Furth 1 and 1 for the Fight Security 1 A Allocatery a variable absorbed of a security of the security absorbed of the security o</pre>	6.4. body. Doration service' many lands a set is high process. The service is a set of the set o

Figure 5

The concordance function comes with its own set of parameters. Most importantly, the user must specify the number of context words (used in the VSM-based part of the experiment) for calculating *relevance*. Theoretically, sentences with higher relevance score tolerate a change from the concordance keyword (in this case: *bank*) to the target word selected in the VSM experiment (here: *river*).

In this example, the following four sentences are returned as the most relevant ones on the basis of the Wikipedia data (relevance scores are specified in parentheses):

- (1) Israel has created roads and checkpoints in the West <u>Bank</u> with the stated purpose of preventing the uninhibited movement of... (1.0)
- (2) Reports IDF PM determined to implement West <u>Bank</u> pullout says Israel not interested in toppling Hamas... (1.0)
- (3) ... Laboratory receives CAP certification for blood <u>bank</u>. (0.99)
- (4) The earliest phase of the Stonehenge monument a circular earth <u>bank</u> and ditch dates to # BC... (0.98)

Bank is a well-known example of homonymy (largely because of its "financial institution" and "river bank" senses). In this particular example, the program was meant to return examples with the "river bank" meaning, since the program was asked to locate sentences in which *river* and *bank* were interchangeable.

Sentences (1) and (2) seem to be correct at first sight. Unfortunately, sentences (1) - (3) are all affected by tokenization issues, the problem of chunking text into appropriate units. *West Bank* in sentence (1) and (2) is a proper name which should have been handled as a single word. The spaceXplorer program received no information about this tokenization option from the corpus (this is a very common problem with multi-word entities). Sentence (3) is similar: *blood bank* is a compound; therefore, it should have been handled as a single lexical item. Sentence (4) is unaffected by these tokenization problems and delivers a sense very

close to the expected "river bank" sense. Please note that the majority of the sentences listed in Figure 5 contain the "financial institution" sense as it is much more frequent (in the Wikipedia subcorpus being used) than the "river bank" sense.

Distributional Semantics collects data about *word types;* characterizing individual tokens (occurrences of a word) is a rarity in the literature. Some notable exceptions are Reisinger and Mooney's (2010) prototype-based approach, Erk and Padó's (2010) exemplars and Scheible, Schulte im Walde and Springorum's (2013) *Codis-Contexts* disambiguation method. They tackled the question of lexical ambiguity (in very different ways) within the framework of Distributional Semantics. The spaceXplorer program offers a traditional, *word type-based* approach to measuring distributional compatibility, without provisions for disambiguity. The concordance function of spaceXplorer lets the user pick out authentic examples of use, while quantifying relevance of each corpus sentence as to the compatibility relation between the selected target word and the selected compatible(s); the user can browse the list and work with the examples that he or she prefers.

As shown above, the proper analysis and interpretation of corpus data continue to require human linguistic knowledge and intuition. In this respect, *distributional concordancing* is not different from traditional concordancing, where frequency, t-score, MI-score etc. information give assistance for the user to find collocations (by creating lists of potential collocations), but it is up to the human user to find the appropriate patterns.

4. Concluding remarks

The spaceXplorer program introduced in this paper is a clean, easily available, free distributional semantic tool, suitable for many situations. The program uses a Wikipedia snapshot and supports linguistic annotation (lemma and POS information). It lets the user carry out corpus-based word similarity experiments and it offers a concordancing facility.

The present form of the program is not suitable for massive qualitative evaluations, compositional distributional studies or for creating large word-context matrices. It does, however, support small and medium-size investigations, qualitative and quantitative studies, run on Windows computers, offer a graphical user interface (GUI) and let the user set the processing parameters easily and see the effect on the screen. It is also designed to be an ideal tool for teaching distributional semantics to students of linguistics and psycholinguistics. The program available for free from the author of is this paper (toth.agoston@arts.unideb.hu).

Acknowledgement: This research was supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP-4.2.4.A/ 2-11/1-2012-0001'National Excellence Program'.

References

Crouch, C. J. 1988. 'A Cluster-based Approach to Thesaurus Construction' in Y. Chiaramella (ed.). SIGIR'88, Proceedings of the 11th Annual International ACM SIGIR Conference. New York: Association for Computational Linguistics, pp. 309-320.

- Erk, K. and S. Padó, 2010. 'Exemplar-based Models for Word Meaning in Context' in J. Hajic, S. Carberry, S. Clark and J. Nivre (eds.). *Proceedings of ACL 2010 Conference Short Papers*. Uppsala: Association for Computational Linguistics, pp. 92-97.
- Manning, C. D., P. Raghavan, H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge, Cambridge University Press.
- Miller, G. A. and W. G. Charles, 1991. 'Contextual Correlates of Semantic Similarity' in Language and Cognitive Processes, vol. 6(1), pp. 1-28.
- Pado, S. and M. Lapata. 2007. 'Dependency-based Construction of Semantic Space Models' in *Computational Linguistics*, vol. 33(2), pp. 161-199.
- Pantel, P. and D. Lin, 2002. 'Discovering Word Senses from Text' in O. R. Zaïane, R. Goebel, D. Hand, D. Keim, R. Ng (eds.). Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, pp. 613-619.
- Pennacchiotti, M., D. D. Cao, R. Basili, D. Croce, M. Roth 2008. 'Automatic Induction of FrameNet Lexical Units' in M. Lapata and H. Tou Ng. (eds.). Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-08), pp. 457-465.
- Reisinger, J. and R. J. Mooney, 2010. 'Multi-Prototype Vector-Space Models of Word VyMeaning' in D. Cer, Stroudsburg, Pennsylvania, Association for Computational Linguistics, C. D. Manning, D. Jurafsky (eds.). *Human Language Technologies: The* 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Proceedings NAACL. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 109-117.
- Sahlgren, M. 2008. 'The Distributional Hypothesis' in *Rivista di Linguistica (Italian Journal of Linguistics)*, vol. 20(1), pp. 33-53.
- Salton, G. 1971. The SMART Retrieval System: Experiments in Automatic Document Processing. Upper Saddle River, New Jersey: Prentice-Hall.
- Scheible, S., S. Schulte im Walde, S. Springorum 2013. 'Uncovering Distributional Differences between Synonyms and Antonyms in a Word Space Model' in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 489-497. [Online]. Available:

http://www.aclweb.org/anthology/I/I13/. [Accessed 2014, May 15].

- Schütze, H. 1998. 'Automatic Word Sense Discrimination' in Computational Linguistics, vol. 24(1), pp. 97-124.
- Tellex, S., B., Katz, J., Lin, A. Fern, G. Marton 2003. 'Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering' in J. Callan, G. Cormack, C. Clarke, D. Hawking, A. Smeaton (eds.). Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Toronto: ACM Press, pp. 41-47.
- Vyas, V. and P. Pantel 2009. 'Semi-automatic Entity Set Refinement' in U. Germann, C. Shah, S. Stoyanchev, C. P. Rose, A. Sarkar (eds.). Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-09, pp. 290-298.
- Widdows, D. 2004. *Geometry and Meaning*. Stanford, California: Center for the Study of Language and Information.