10.12871/CLICIT2014138

ConParoleTue: crowdsourcing al servizio di un *Dizionario delle Collocazioni Italiane per Apprendenti* (*Dici-A*)

Andrea Gobbi
Dipartimento di Scienze Politiche, Sociali e della Comunicazione, Università di Salerno
andgobbi@gmail.com

Stefania Spina
Dipartimento di Scienze Umane e Sociali, Università per Stranieri di Perugia
stefania.spina@unistrapg.it

Abstract

English. *ConParoleTue* è un esperimento di uso del crowdsourcing nell'ambito della lessicografia L2. A partire dalla costituzione di un dizionario di collocazioni per apprendenti di italiano L2, *ConParoleTue* rappresenta un tentativo di re-inquadramento di problematiche tipiche dell'elaborazione lessicografica (la qualità e il registro delle definizioni) verso una maggiore centralità delle necessità comunicative di chi apprende. A questo fine una metodologia basata sul crowdsourcing viene sfruttata per la redazione delle definizioni. Questo articolo descrive tale metodologia e presenta una prima valutazione dei suoi risultati: le definizioni ottenute attraverso il crowdsourcing sono quantitativamente rilevanti e qualitativamente adatte a parlanti non nativi dell'italiano.

Italiano. *ConParoleTue* is an experiment of adoption of crowdsourcing techniques applied to L2 lexicography. It started while compiling a dictionary of collocations for learners of Italian as a second language, and it uses crowdsourcing to find new solutions, both quantitatively and qualitatively, to traditional issues connected with lexicography, such as the quality and the register of definitions, towards a more learner-centred approach. This paper describes our methodology and a first evaluation of results: the definitions acquired through crowdsourcing are quantitatively relevant and qualitatively appropriate to non-native speakers of Italian.

1 Introduzione

ConParoleTue (2012) è un esperimento di applicazione del crowdsourcing all'ambito della lessicografia L2, elaborato all'interno del Progetto APRIL (Spina, 2010b) dell'Università per Stranieri di Perugia nel corso della costituzione di un dizionario di collocazioni per apprendenti di italiano L2.

Le collocazioni occupano da alcuni decenni un posto di primo piano negli studi sull'apprendimento di una lingua seconda (Meunier e Granger, 2008). Quella collocazionale è riconosciuta come una competenza chiave per un apprendente, perché svolge un ruolo fondamentale nei due aspetti della produzione (fornisce infatti blocchi lessicali precostituiti e pronti per essere utilizzati, migliorando la fluenza; Schmitt, 2004) e della comprensione (Lewis, 2000). Anche nell'ambito della lessicografia italiana la ricerca sulle collocazioni è stata particolarmente produttiva, ed ha portato, negli ultimi cinque anni, alla pubblicazione di almeno tre dizionari cartacei delle collocazioni italiane: Urzì (2009), nato in ambito traduttivo; Tiberii (2012) e Lo Cascio (2013).

Il *DICI-A* (*Dizionario delle Collocazioni Italiane per Apprendenti*; Spina, 2010a; 2010b) è costituito dalle 11.400 collocazioni italiane estratte dal *Perugia Corpus*, un corpus di riferimento dell'italiano scritto e parlato contemporaneo[1].Tra le tante proposte, la definizione alla base della costituzione del *DICI-A* è quella di Evert (2005), secondo cui una collocazione è "a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon". Le collocazioni del *DICI-A* appartengono a 9 categorie diverse, selezionate sulla base delle sequenze più produttive di categorie grammaticali che le compongono: aggettivo-nome (*tragico errore*), nome-aggettivo (*anno prossimo*), nome-nome (*peso forma*), verbo-(art.)-nome (*fare una domanda/fare pena*), nome-preposizione-nome (*carta di credito*), aggettivo-*come*-nome (*fresco come una rosa*), aggettivo-congiunzione-aggettivo (*sano e salvo*), nome-congiunzione-nome (*carta e penna*), verbo-aggettivo (*costare caro*).

Per ogni collocazione sono stati calcolati gli indici di Juilland di dispersione e di uso (Bortoli-

[1] http://perugiacorpus.unistrapg.it

ni et al., 1971), sulla base dei quali sono state selezionate le collocazioni definitive. Si è presentato dunque il problema di come procedere alla loro definizione. In questo contesto è nata l'idea dell'impiego del crowdsourcing, e all'elaborazione di *ConParoleTue*.

2 La scelta del crowdsourcing

L'adozione del crowdsourcing in linguistica è principalmente legata ad obiettivi di ottimizzazione delle risorse (Snow et al., 2008; Hsueh et al., 2009), in particolare nell'ambito della traduzione (Callison-Burch, 2009), della creazione di corpora (Wang et al., 2012; Post et al.,2012) e della loro annotazione (Munro et al., 2010); tra le metodologie e gli strumenti più utilizzati figurano *Mechanical Turk* di Amazon (Schnoebelen e Kuperman, 2010) e i serious games (Kneissl e Bry, 2012).

Oltre all'aspetto dell'ottimizzazione delle risorse, tuttavia, la scelta del crowdsourcing per il *DICI-A* è stata dettata anche da un preciso approccio alla lingua,che presta particolare attenzione alla natura sociale e condivisa dello strumento linguistico, da cui derivano i suoi specifici processi acquisizionali (Gobbi, 2012; Gobbi, 2013; Gobbi e Spina, 2013).

Il coinvolgimento di una platea molto ampia di collaboratori per acquisire le definizioni delle collocazioni da includere nel dizionario, e il modo stesso con il quale il progetto è stato presentato (ogni richiesta di definizione recitava: "Come lo spiegheresti ad un tuo amico straniero?") era volutamente teso ad elicitare il maggior grado possibile di naturalezza e spontaneità nelle risposte. Da un punto di vista meta lessicografico, ciò ha comportato la decisione di non richiedere ai contributori di conformarsi ad uno stile predeterminato di definizione, allo scopo di perseguire le condizioni di informalità dell'interazione quotidiana. I vantaggi di un tale approccio collaborativo, sviluppato dal basso e mirato alla naturalezza delle definizioni, sono diversi, e di diversa natura: in primo luogo, quello di offrire agli apprendenti e futuri utenti del *DICI-A* uno strumento che fornisca risposte meno accademiche e più formalmente simili a quelle ottenibili nella vita quotidiana, e dunque adeguate ad un contesto interazionale. Un tale approccio, inoltre, si presta alla sensibilizzazione di parlanti nativi su questioni linguistiche, quali il dover riflettere su come definire un'espressione con altre parole, operazione di fatto non semplice (Schafroth, 2011). Infine, lo sviluppo di uno strumento di riferimento per apprendenti di una L2 come un'opera collettiva, sebbene monitorato e revisionato nella sua forma finale, rappresenta una sfida interessante ed ambiziosa, oltre che un esperimento applicativo di metodologie che sempre più spesso si rivelano preziose nella ricerca linguistica.

2.1 Metodologia

Per la realizzazione dell'esperimento, è stata innanzitutto predisposta una piattaforma web dedicata[2]. Dopo una breve schermata di presentazione, attraverso la piattaforma vengono raccolti pochi dati essenziali sui partecipanti (età, sesso, titolo di studio, madrelingua, eventuale livello QCER di italiano), al fine di acquisire alcune informazioni sociolinguistiche di base su ciascuno degli autori delle definizioni.

Il sistema propone quindi, una dopo l'altra, cinque collocazioni da definire, estratte a caso dal database (fig. 1).

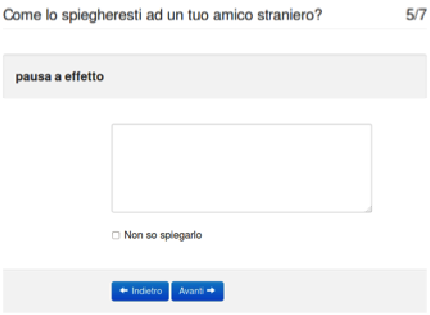

Figura 1 - Esempio di collocazione da definire

Infine, viene chiesto di valutare due definizioni precedentemente elaborate da altri partecipanti, allo scopo di effettuare un primo filtraggio di eventuali definizioni inappropriate (fig. 2).

Il progetto è stato ampiamente diffuso tramite vari social media (una pagina Facebook e un profilo Twitter appositamente creati), una newsletter dedicata, e numerosi contatti istituzionali. Sebbene sia tuttora online, il periodo di maggior attività di *ConParoleTue* è stato quello compreso tra dicembre 2012 ed aprile 2013, data di conclusione del progetto APRIL.

[2]http://elearning.unistrapg.it/conparoletue/

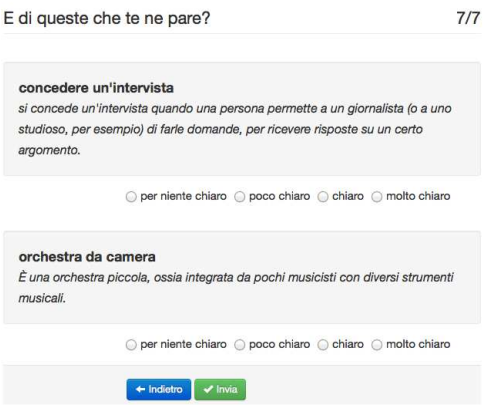

Figura 2 - Esempio di valutazione delle definizioni

3 Risultati

Le definizioni ottenute attraverso l'esperimento di crowdsourcing erano, a marzo 2014, 3.267 (al netto di una ventina redatte in lingue diverse dall'italiano, e di poche altre illeggibili). Per verificare le caratteristiche di tali definizioni, elaborate non da specialisti, ma da semplici parlanti dell'italiano, esse sono state confrontate con un numero identico di definizioni tratte da un dizionario monolingue, il De Mauro Paravia (2000); le 3.267 definizioni del De Mauro sono state estratte in modo casuale tra quelle riferite a una sola delle possibili diverse accezioni di lemmi di marca comune. Il confronto con le definizioni elaborate da lessicografi mira a verificare l'ipotesi di una maggiore naturalezza delle definizioni create da parlanti non specialisti e, di conseguenza, della loro appropriatezza per un dizionario delle collocazioni destinato a parlanti non nativi dell'italiano. Tra le caratteristiche principali di un *learner dictionary*, che ne fanno uno strumento anche concettualmente diverso rispetto ad un dizionario per parlanti nativi (Tarp, 2009), c'è infatti proprio la specificità delle sue definizioni: in quanto rivolte ad un pubblico di parlanti non nativi, esse dovrebbero:

- avere carattere più linguistico che enciclopedico, quindi "evocare un tipo di sapere pre-scientifico, intuitivo, [...] che abbia un valore prototipico, facilmente riconoscibile" (Schafroth, 2011:26);
- essere formate da un lessico semplice, per quanto possibile di base, e da una sintassi poco complessa, adatta alle limitate competenze linguistiche dei destinatari.

Un *learner dictionary* dovrebbe far comprendere ai lettori il significato di un'espressione facendo riferimento quanto più possibile a conoscenza generica e condivisa e non caratteristica della lingua target, fornendo loro il maggior numero di informazioni possibile sui suoi contesti sintagmatici (Schafroth, 2011).

La presenza di queste caratteristiche può essere verificata attraverso alcune misure quantitative calcolate nel corpus di definizioni; nel confronto tra quelle ottenute attraverso l'esperimento di *ConParoleTue* (d'ora in avanti CPT) e quelle del dizionario De Mauro (DM) abbiamo dunque considerato in primo luogo aspetti superficiali dei due testi, come il numero di tokens per definizione e la lunghezza media delle parole, aspetti tradizionalmente associati alla maggiore o minore semplicità di un testo (Franchina e Vacca, 1986). I risultati, riassunti nella tab. 1, mostrano come le definizioni di CPT siano più brevi di quelle di DM,mediamente composte da parole più brevi e da un numero maggiore di frasi più brevi.

	tokens	tokens per definizione	frasi	tokens per frase	lunghezza parole
CPT	38.697	11,8	3.506	11,2	5
DM	42.310	13,2	3.318	13	5,7

Tabella 1 - Misure quantitative di CPT e DM

I tratti superficiali considerati fin qui sono quelli che tradizionalmente concorrono al calcolo dell'indice di leggibilità (Amizzoni e Mastidoro, 1993), che ha appunto l'obiettivo di misurare il grado di facilità con cui un testo viene letto e compreso; uno degli indici di leggibilità più utilizzati per l'italiano, *Gulpease* (Lucisano e Piemontese, 1988), differisce in modo significativo in CPT (68,7) e DM (60,59).

Se tutti questi elementi suggeriscono una maggiore comprensibilità delle definizioni ottenute attraverso il crowdsourcing, vanno comunque considerati i limiti degli indici, che, come quello di *Gulpease*, sono basati esclusivamente su caratteristiche superficiali dei testi, come la lunghezza in caratteri delle parole e quella delle frasi; tali caratteristiche hanno dimostrato di essere indicatori spesso non del tutto attendibili della leggibilità dei testi (vedi ad esempio Feng et al., 2009).

Per valutare in modo più accurato il grado di comprensibilità dei due gruppi di definizioni, in particolare per parlanti non nativi dell'italiano, abbiamo considerato una serie di altri tratti, di tipo lessicale e morfosintattico (Heilman et al., 2007), sulla base di alcune delle indicazioni contenute in Dell'Orletta et al., (2011).

I tratti lessicali comprendono il rapporto tra types e tokens (TTR), che misura la varietà del lessico utilizzato, e la distribuzione dei tokens di CPT e DM nelle tre fasce di frequenza del vocabolario di base. La TTR[3], considerato uno degli indicatori della leggibilità di un testo (Dell'Orletta et al.,2011), è risultata significativamente più elevata in DM (49,4) rispetto a CPT (36.3).

Per misurare la distribuzione dei lemmi delle definizioni nelle tre fasce del vocabolario di base è stata utilizzata la lista di frequenza dei lemmi estratti dal *Perugia Corpus*; in particolare, la fascia dei 2000 lemmi più frequenti (rango 1-2000), che copre il 79% dei lemmi totali del corpus, la fascia dei successivi 2000 lemmi (rango 2001-4000), che aggiunge alla precedente una copertura del 5,9%, e la fascia dei successivi 3000 lemmi (rango 4001-7000), che aggiunge una copertura del 3,4% dei lemmi totali. Le tre fasce, dunque, comprendono i 7000 lemmi più frequenti del *Perugia Corpus*, che totalizzano una copertura dell'88,3% e che sono assunti come vocabolario di base[4]. La fig. 3 rappresenta la diversa distribuzione dei lemmi delle definizioni nelle tre fasce di frequenza; il grafico evidenzia come in CPT siano predominanti i lemmi della fascia più frequente, quindi quelli più verosimilmente già noti a parlanti non nativi di italiano, mentre in DM oltre il 20% dei lemmi è composto da parole non incluse tra le 7000 più frequenti, e in particolare da nomi astratti o poco comuni (*intasamento*, *lamina*, *perno* o *merlatura).*

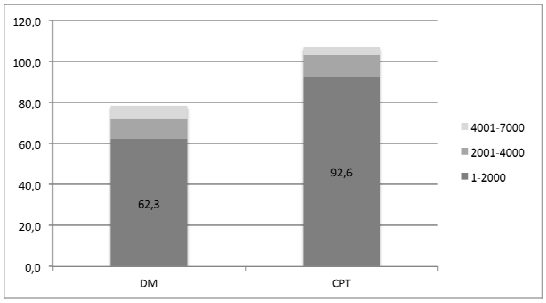

Fig. 3 - La distribuzione dei lemmi di DM e CPT nelle tre fasce del vocabolario di base

Passando infine agli aspetti morfosintattici, nei due corpora di definizioni sono stati misurati i tratti riportati nella tab. 2 (i verbi, i nomi, e tre tipi di frasi subordinate: quelle implicite introdotte da preposizioni, quelle esplicite introdotte da congiunzioni, e le relative). Per ognuno dei tratti è stato calcolata la log-likelihood (Rayson e Garside, 2000), per misurare la significatività delle differenze. Come si evince dalla tab. 2, le definizioni di CPT sono composte da un numero sensibilmente maggiore di verbi (specie di modo finito e per il 90% inclusi nei 2000 lemmi più frequenti) e da un numero minore di nomi; CPT si serve inoltre in misura significativamente maggiore di subordinate, sia implicite che esplicite. Come mostra la coppia di esempi (1) e (2), le definizioni non specialistiche di CPT procedono per brevi subordinate che precisano con parole semplici l'enunciazione della principale, mentre quelle di DM, spesso prive di verbo,sono caratterizzate da un accumulo di sintagmi nominali e preposizionali, per lo più astratti.

(1) Pietra dello scandalo (CPT): qualcuno che è al centro dell'attenzione perché ha fatto qualcosa di grave.

(2) Scandalo (DM): turbamento della coscienza o sconvolgimento della sensibilità.

Tratto	*CPT*	*DM*	*L-L*	*p-value*
Verbi	6185	5746	79,10	0,000
Nomi	8525	9803	11,61	0,001
pre. + sub.	849	388	219,63	0,000
cong. sub.	1516	699	385,92	0,000
rela. ≠CHE	257	183	19,99	0,000

Tabella 2 - Tratti morfosintattici in CPT e DM

4 Conclusioni

L'esperimento descritto, che riguarda l'uso del crowdsourcing per l'acquisizione di definizioni di collocazioni italiane redatte da parlanti generici, si è rivelato efficace sia dal punto di vista quantitativo (oltre 3200 definizioni raccolte in cinque mesi) che da quello della loro appropriatezza ad un pubblico di apprendenti. Un confronto con definizioni redatte da un team di lessicografi ha evidenziato il carattere più intuitivo e naturale delle definizioni dei non specialisti, rispetto alla maggiore astrattezza e complessità delle definizioni dei professionisti. I risultati descritti inducono a proseguire la redazione del dizionario attraverso tale metodologia basata sul crowdsourcing.

[3] La TTR è stata calcolata usando l'indice di Guiraud (Guiraud, 1954), per ovviare alla non omogeneità nel numero dei tokens dei due insiemi di dati.

[4] Il *Vocabolario di Base della lingua italiana* (De Mauro 1980) è in corso di revisione. Per questo si è deciso di utilizzare al posto della sua vecchia versione la lista di frequenza dei lemmi del *Perugia Corpus*, anche se non rappresenta nativamente un vocabolario di base dell'italiano.

References

Maurizio Amizzoni e Nicola Mastidoro. 1993. Linguistica applicata alla leggibilità: considerazioni teoriche e applicazioni. *Bollettino della Società Filosofica Italiana*, n. 149 (maggio - agosto 1993), pp. 49-63.

Umberta Bortolini, Carlo Tagliavini e Antonio Zampolli. 1971. *Lessico di frequenza della lingua italiana contemporanea*. Garzanti, Milano.

Chris Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 286–295.

ConParoleTue. 2012. Home Page del progetto: http://elearning.unistrapg.it/conparoletue.

Tullio De Mauro. 1980. *Guida all'uso delle parole*. Editori Riuniti, Roma.

Tullio De Mauro. 2000. *Dizionario della lingua italiana*. Paravia, Torino.

Stefen Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* Dissertation, IMS, University of Stuttgart.

Lijun Feng, Noemie Elhadad and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (EACL '09), pp. 229–237

Valerio Franchina e Roberto Vacca. 1986. Adaptation of Flesh readability index on a bilingual text written by the same author both in Italian and English languages. *Linguaggi* (3), pp. 47-49

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pp. 27–35.

Fabian Kneissl and François Bry. 2012. MetropolItalia: a crowdsourcing platform for linguistic field research. *Proceedings of the IADIS international conference WWW/internet*.

Andrea Gobbi. 2012. Ipotesi Glottodidattica 2.0. *Journal of e-Learning and Knowledge Society*, 8(3): 47-56.

Andrea Gobbi. 2013. Tweetaliano: a native 2.0 approach to language learning. *ICT for Language Learning 2013, Conference Proceedings*, 282-285.

Andrea Gobbi e Stefania Spina. 2013. Smart Cities and Languages: The Language Network. *Interaction Design and Architecture(s) Journal – IxD&A*. 16: 37-46.

Paul Guiraud. 1954. *Les Charactères Statistiques du Vocabulaire. Essai de méthodologie*. Presses Universitaires de France, Paris.

Michael J. Heilman, Kevyn Collins and Jamie Callan. 2007. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. *Proceedings of the Human Language Technology Conference*, pp. 460–467

Michael Lewis. 2000. *Teaching collocation. Further developments in the lexical approach*. Language Teaching Publications, Hove.

Vincenzo Lo Cascio. 2013. *Dizionario Combinatorio Italiano*. John Benjamins, Amsterdam.

Pietro Lucisano e Maria Emanuela Piemontese. 1988. GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana, *Scuola e città*, 3, 31, marzo 1988, pp. 110-124.

Fanny Meunier e Sylviane Granger. 2008. *Phraseology in foreign language learning and teaching*. John Benjamins, Amsterdam.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 122-130.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for computational linguistics, 401-409.

Progetto April. 2010. Home Page del progetto: http://april.unistrapg.it/april/.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. *Proceedings of the workshop on Comparing Corpora*, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000). 1-8 October 2000, Hong Kong, pp. 1 - 6.

Elmar Schafroth. 2011. Caratteristiche fondamentali di un learner's dictionary italiano. *Italiano Lingua Due*, 1, pp. 23-52.

Norbert Schmitt (Ed.). 2004. *Formulaic Sequences*. John Benjamins, Amsterdam.

Tyler Schnoebelen and Victor Kuperman. 2010.Using Amazon Mechanical Turk for linguistic research. *Psihologija*, Vol. 43 (4), 441–464.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew T. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natu-

ral language tasks. *EMNLP '08: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263.

Stefania Spina. 2010a. The Dici Project: towards a Dictionary of Italian Collocations integrated with an online language learning platform, in Granger S., Paquot M., *eLexicography in the 21st century: New Challenges, New Applications*, Proceeding of eLex 2009 (Louvain-La-Neuve, 22-24 ottobre 2009), Presses Universitaires de Louvain, pp. 273-282.

Stefania Spina. 2010b. The Dictionary of Italian Collocations: Design and Integration in an Online Learning Environment, in Calzolari N., Choukri K., Maegaard B., Mariani J., Odjik J., Piperidis S., Rosner M.and Tapias D., 2010, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Malta, May 2010, European Language Resources Association, pp. 3202-3208 .

Sven Tarp. 2009. The foundations of a theory of learners' dictionaries. In *Lexicographica*, 25, pp. 155-168.

Paola Tiberii. 2012. *Dizionario delle collocazioni*. Zanichelli, Bologna.

Francesco Urzì. 2009. *Dizionario delle Combinazioni Lessicali*. Convivium, Lussemburgo.

William Yang Wang, Dan Bohus, Ece Kamar and Eric Horvitz. 2012. Crowdsourcing the acquisition of natural language corpora: Methods and observations. *Proceedings of the IEEE SLT 2012*, 73-78.