Linguistic resources and MT trends for the Italian language: overview and perspectives

Isabella Chiari
Dipartimento di scienze documentarie, linguistico-filologiche e geografiche
Sapienza Università di Roma
P.le Aldo Moro ,5
00185 Roma, Italy
isabella.chiari@uniroma1.it

Abstract — This contribution aims at providing an overview of recent trends in the integration and development of linguistic resources for the Italian language to be used in machine translation and other computational tools. After a brief outline of the main perspectives and trends in MT, we will specifically address questions relating to parallel and comparable corpus building and integration, thesauri and WordNets development, and knowledge bases for Italian and its main challenges.

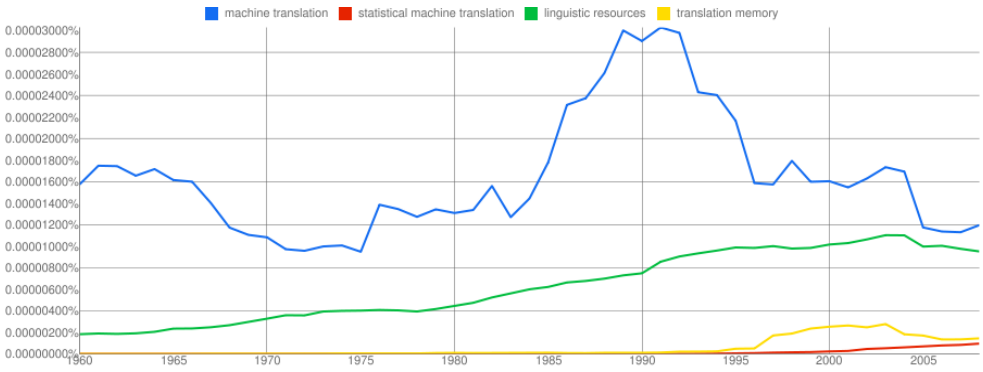
Keywords — machine translation, computational linguistics, linguistic resources, knowledge bases, Italian language

Introduction

Translation automation whether it is conceived as machine translation, computer-aided translation or human-aided machine translation has undergone significant changes in the last decades and has proven to be effective in respect to specific objectives. In the last decade we have witnessed a large interest in previously underrated languages, such as Italian, that is in of traditionally peripheral interest in computational linguistics and disposes of far less tools and applications, especially in the area of open source projects.

Linguistics has provided different waves of innovation in the technological field and in the theoretical background necessary to develop complex integration tasks among available resources. Italian linguistics is no exception [1, 2].

Fig. 1 Google Books *N*gram Viewer on key terms (English dataset)



Just to give an idea, and with the necessary caution, we can observe trends in the Google Books Ngram Viewer (Fig. 1): while the destiny of the expression "machine translation" *tout court* has been controversial over the past fifty years, the notion of linguistic resource is gathering more and more space in the literature. It must be noted that statistical data for the last decade on *N*gram Viewer are not reliable for sampling reasons [3], nevertheless even if we focus to data up to the year 2000 the general tendency is confirmed.

Trends in MT: corpora and the web

Aims and objectives of MT have significantly changed mainly directed towards different methodologies for general purpose translation and controlled language translation. Statistical and Corpus based MT has proven to be an effective and economical alternative to the traditional rule-based MT, given the large availability of electronic texts of general and special domains. Data driven MT is tightly connected to corpus research both parallel and multilingual. A clear sign of this trend is Google's conversion to SMT and results of evaluation tasks that indicate good comparative results for n-gram based MT systems ([4]).

A second point is the assumption, widely observed in manual and automatic evaluation, that MT systems do not work successfully at the same level with different language pairs [5, 6], depending on different language and feature similarities and differences. Different models of SMT (word-based, phrase-based or syntax-based) seem to be more effective depending on specific morphological and syntactic features of the single languages that enter a pair. Furthermore performance of machine translated texts varies considerably depending on the translation direction (from or into a certain language).

Thus capital is the availability of monolingual and bilingual corpora and of tools for sentence-alignment for the construction of parallel corpora. Availability is often constrained to specific language pairs and quality strongly depends on the amplitude and quality of textual data (as easily seen on LDC corpora, http://www.ldc.upenn.edu, or Europarl corpus, http://www.statmt.org/europarl/). Same applies to monolingual corpora used for language modelling.

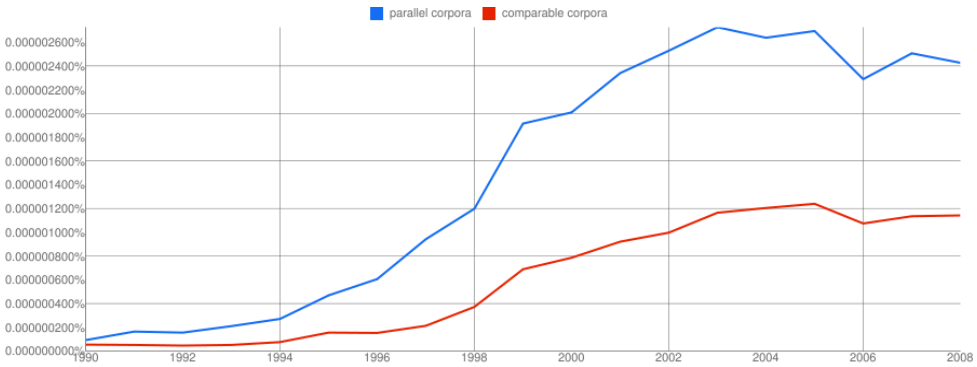
The use of parallel corpora as training corpora for SMT (Statistical Machine Translation) can give very good results when the domain is fairly specific and when the extension of the corpus is about 30-40 million words. In these cases SMT system can obtain competitive performance over commercial systems [7, 8]. Otherwise good results can be obtained with small training sets on a specific domain [9]. But parallel corpora are very difficult to build because of limitations in size, coverage and register.

The "more data are better data" advice of Church and Mercer [10] is to be taken carefully because, as Philip Resnik [11] points out, since it cannot guarantee homogeneity of linguistic and semantic features of texts and it can distort application outputs.

Comparable corpora can in some way be a shortcut solution for some purposes. In some cases small parallel corpora and larger comparable corpora can be jointly used in order to train machine translation systems (e.g. the Muntaneau & Marcu procedure [12]). Still in favour of the use of comparable corpora is the observation that "a domain-specific comparable Web corpus provides better translations than a general-purpose parallel corpus, even if the latter is of much higher alignment quality" [13], even if it is generally acknowledged that "Web-mined parallel corpus, despite its smaller size, improves SMT much more than Web-mined comparable corpus" [14].

An idea on the general interest in parallel corpora vs comparable corpora can be seen again in the Google *N*gram viewer Fig. 2, where we observe a fast growth of comparable corpora, especially in the last decade, where parallel corpora have a small shift downwards.

Fig. 2 Google *N*gram Viewer for parallel and comparable corpora (English dataset)



In order to partially solve these problems, tools for the automatic extraction of parallel corpora from the web are being developed ([9, 15-19]). Methods for detecting parallel texts are, for example, filename and path similarity comparison, anchor names and links, file makeup, layout or design comparison, and content-based similarity evaluation. An overview of the main differences in features used to perform this task is presented in Table 1.

Resnik's strand (Structural Translation Recognition, Acquiring Natural Data) [11, 17] for example mainly relies on the presence of names of languages in hyperlinks or URLs in webpages, while other methods use annotated and machine translated texts to enlarge training corpora for SMT [20]. Strand is based "on the insight that translated Web pages tend quite strongly to exhibit parallel *structure*, allowing them to be identified even without looking at content" [17], while more recent version use an integrated method that includes content analysis.

Table 1 Deriving parallel corpora from the web

Methodology	Main features	Authors
STRAND	• Anchors and links • Language identification • Structural recognition algorithm	Resnik (1998; 1999)
BITS	• Domain filtering • Language identification • Content based translation identification	Ma & Liberman (1999)
PTMiner	• Anchors • Host names	Chen & Nie (2000)
PTI	• Filename comparison • Content analysis	Chen et al. (2004)
Babylon	• Seed text • URL and other similarity checks • Tailored on low density languages	Mohler & Mihalcea (2008)
GWB	• Bootcat-like • Seed words • Domain filter	Almeida (2010)

Ma & Liberman's bits methodology (Bilingual Internet Text Search) [21] use language identification and content filtering, applied to an initial domain filtering (depending on the fact that e.g. 1 out of 100 websites in the .de domain possess a parallel English version, while the opposite in not true). Chen & Nie [22] also propose the PTMiner (Parallel Text Miner) system for automatic deriving parallel corpora from the web, a system that not only uses anchors such as in Resknik's algorithm, but also host names and the intersection of the two distinct sets of sites in the two different languages. Chen et al. [19] developed PTI (Parallel Text Identification System) which combines filename comparison with content analysis.

The Babylon Parallel Text Builder [9] is specifically addressed towards developing tools for low density languages (languages scarcely represented on the net). GWB (GetWebBitext) [15] is a fully automatic tool for the extraction and parallel corpus processing from the web using seed words in the tradition of Bootcat and WebBootcat (e.g. [23]).

Among those methods there are emergent procedures that use only textual mining techniques [14], adapting existing techniques to an open set of documents (the Web) and focusing on document filtering and sentence alignment. The usefulness of these procedure is more effective when it can be applied to an open set of document of which we do not need access to metadata, we do not need previous annotation or cleaning and we do not already know in advance whether they contain parallel or comparable texts. It is of course possible to apply those methods to a restricted closed set of documents with even better results, but the efficacy of the procedures and its extendibility lies in its being open and not language-specific.

Hong et al. [14] propose a wider perspective with the aid of search engines in the initial phase of document retrieving, with keyword ranking and representative ranking for further sentence pair extraction.

Some experiments in web extraction of comparable corpora are being conducted extending some of the approaches above mentioned, even if the task is far from simple and needs to be focused on specific domains [13, 24]. While parallel corpus extraction needs the definition of severe parameters to restrict the candidates to real translations, comparable corpora extraction need looser pairing strategies, but risks noisy document extraction. Talvensaari*, et al.* focus on domain vocabularies that trigger a process of URL selection and relevance (according to the semi-manually built domain vocabulary of keywords).

While identification of translational patterns from parallel corpora is now widely spread, more interesting seems the approach that generates automatic translational patterns from unrelated corpora in different languages (e.g. [25] applied to English and German; [12, 26, 27]).

Another very promising approach is the use of Wikipedia as a source of comparable corpora, whether in full form using corresponding pages in different languages, whether using the Wikipedia structure to extract parallel terminology (e.g. [28-33]). Last but not least the

approach in Nagata et al. [34] on the exploitation of partially bilingual texts (where in a L1 text are inserted pieces of terminology in L2) to build bilingual lexicons seems also to be promising.

So from the methodological point of view in commercial as well as in academic research it is nowadays unavoidable the use of corpora, whether parallel or comparable, in order to make advances in all fields of machine translation and NLP in its wider sense. Even though in the last decade several researchers have proposed different techniques for automatically extracting parallel and comparable corpora from the web, still the major problems of representing minor languages, of gathering data on different domain specific fields or textual typologies are solved in ad hoc procedures often deliberately constrained to specific research projects and aims and not fully extendable to other domains or purposes.

The great variability of information on web pages currently on the net, the variety of encodings (of data and metadata) make it very difficult to elaborate a long term strategy in this direction (even though in some cases standard content management systems include language codes that can be profitably used to extract multilingual data, though this is just a small part of the problem).

There is a great demand for tools flexible enough to be used for small and large corpus collection and that can be tailored to specific user needs, but that exhibit transparency in procedures and full control over technical and methodological decisions.

Linguistic resources for Italian

In the Nineties the notion of *linguistic database* has gained great dignity in the linguistic field, where the term database is intended as "collections of electronic records of linguistic data" [35], shifting progressively towards the notion of *linguistic resource* that integrate different kinds of linguistic data with encyclopaedic information in the form of knowledge bases. Furthermore the notion of linguistic resource is thought of more as a multiuse storage of data to be integrated within different computational tools, not only to be queried autonomously.

Parallel and comparable corpora are widely used in NLP (word sense disambiguation, anaphora resolution, information extraction and retrieval, cross-language information retrieval, etc.) research and, specifically, for the training of SMT. Availability of parallel or comparable corpora heavily depends on text typology already accessible on the web in digital form. The source of these texts has often been institutional, being produced mainly by organizations as the European Union or the United Nations. An example for all is the Europarl corpus of proceedings of the European Parliament and containing about 50 million words for each of the 21 official languages (from 1996 and updated up to the 2010 proceedings). Other typical domains that are relatively easily available is those of religious texts [36] and localized software instruction manuals [37]

The need of parallel corpora to be used in translation tools for subject different from law, medicine, politics and specific domains is more and more explicit, even when digital texts are widely available on the net., because the web does not equally represent all text and language typologies.

For the normative domain EU documents form a large base of data for the related languages. Italian is well represented only in the European Corpus Initiative Multilingual Corpus I (ECI/MCI), the Europarl corpus, containing 49,981,015 tokens for the Italian-English parallel version [6] and in the JRC-ACQUIS parallel corpus of European Commission regulations [38, 39] (ca. 50 million words for each EU language; 57 million words for Italian), and fairly represented in the OPUS parallel corpus of technical and normative texts [40, 41].

Another Italian parallel corpus is the MultiSemCor corpus (Italian-English, http://multisemcor.fbk.eu/index.php) created from a subset of 200,000 running words from the Brown Corpus. For the fictional text typology we dispose of the CEXI bidirectional corpus of Italian and English texts and translations [42, 43], which is a small scale corpus developed by the School for Interpreters and Translators at the University of Bologna at Forlì (about 500,000 words for each language, a small portion of which is non-fictional); the TEC corpus (Translational English Corpus http://www.llc.manchester.ac.uk/ctis/research/english-corpus) developed by the University of Manchester, containing translations from Italian (and many other languages) of mainly novels and fictional

texts into English.[3] The university of Turin [44] has developed the NUNC Corpus (Newsgroup Usenet Corpora) that can be considered comparable corpora in computer-mediated Communication.

But still many domains are not covered. When we think about the possible exploitation of parallel and comparable corpora of Italian, we still perceive that there is a long way towards us. Probably the application of some of the afore mentioned web mining systems might help in the development and distribution of large domain-specific parallel corpora to be made available in open source mode for research and application.

The availability of monolingual corpora is equally capital since it can be used for language modelling and for annotation training. Standards in the size of monolingual reference corpora have grown from the golden standard of the 100 million words of BNC (British National Corpus) to significantly larger extensions. For English the best balanced and most interesting corpus for linguistic research purposes seems to be the Bank of English (now at 450 million words), although for size we now dispose of definitely larger web corpora. Examples are the corpora extracted from the web of the Wacky series [45-47], available through the Sketch Engine, whose sizes are presented in Table II.

As we can see, Italian has one of the largest corpora, even though the Wacky corpora still contain some duplicates and noise and are not fully and controllably balanced.

Table 2 Monolingual Corpora derived from the Web

Corpus	Language	Words
EnTenTen	English	2,759,126,991
ukWac	English	1,318,047,961
frWac	French	1,279,937,839
deWac	German	1,336,258,089
itWac	Italian	1,575,489,232

Google Books Ngram (361 billion English words) and Web Ngram and Microsoft Ngrams are often used as substitutes for monolingual corpora because of their size [3, 48], but unfortunately they do not contain an Italian data set (yet).

[3] Further parallel resources for languages other than Italian can be found indexed in http://home.sslmit.unibo.it/corpora/alf_index.php.

Among other kinds of relevant linguistic resources to be integrated in different translation tools are thesauri and wordnets. The well-known experience of Wordnet [49] has been pivotal in many computational linguistic fields. Italian has seen some competing projects in this area [50-52]. For example, ItalWordnet (developed in ILC Pisa, within EuroWordNet (EWN) and SI-TAL) [53, 54] and now containing 47,000 lemmas, 50,000 synsets e 130,000 semantic relations. Multiwordnet [55], developed in ITC-IRST Trento, also comprises an Italian WordNet, which contains about 58,000 Italian word senses and 41,500 lemmas organized into 32,700 synsets aligned whenever possible with Princeton WordNet English synsets. There is also a project for the development of an Italian FrameNet [56] using automatic data extraction from corpora.

A different approach is that of LexIt, an online database developed by A. Lenci at the Laboratory for Computational Linguistics of the University of Pisa, in which argument structures of Italian verbs and nouns (with syntactic slots, lexical sets, semantic classes) are presented connected with corpus based data (from Repubblica and Wikipedia).

Methodologies for the semi-automatic construction of resources that integrate dictionary-like resources with lexical knowledge bases (LKB) have been developed and experimented for Italian using English language resources as bridges [52]. The underlying assumption is that English-based synsets can be largely re-used for the development of similar resources for other Indo-European languages. This has not proven to be always true, especially when dealing with languages with diverse morphological richness.

Further directions go in the sense of the integration of lexical resources with ontologies and with alignment of available resources for different languages. Ontologies serve as core of knowledge bases that can be used for the enhancement of translation systems giving birth to (with rather out-dated expression) Knowledge Based Machine Translation (KBMT) [57-59]. Knowledge bases to be used in machine translation are actually quite heterogeneous objects, they may merge dictionaries, ontologies, wordnets and any kind of linguistic or information resources and should be able to analyse linguistic and encyclopaedic meanings.

The enrichment of ontological information on thesauri such as WordNet has proven to be a very useful activity since it makes the thesaurus capable of being used for more complex tasks of automatic reasoning and other high level computational tasks (as automatically extract semantic relations useful for the enrichment of semantic nets [60]).

One example is the open source project Senso Comune [61-63] for the creation of a lexical knowledge base for Italian. In this resource every word meaning of the most common 2,000 nouns of the Italian language has been (manually) associated to a top level ontology (Chiari et al. [64]), while other projects are developing automatic tools to perform alignment among this kind of resources [65-67]. Specific applications for translation terminology automation are developed for Italian in Sapienza (Engineering) by Navigli & Velardi [68], in conjunction with automatic development of glossaries [69].

Projects in integrating (Italian) Wikipedia (encyclopedic and linguistic knowledge) with automatically formalized ontology mapping are being developed in CNR – Rome.

Conclusions

As we have seen, the last decade has observed a dominating trend toward the integration of different methodologies and resources and has also witnessed a general trend in the exploitation of the web to acquire monolanguage and multilanguage data.

The use of internet data cannot be assumed uncritically since it poses many problems of representativeness (of domains, textual typologies and, most of all, of different languages) and contains much noise. Success of SMT using very large training sets, such as that of the Google Translate engine, has posed many questions on the relationship between quantity and quality of data. Is it really true that "more data is better data"? Much of recent research has focused on smaller sets of very coherent and domain specific corpora obtaining good results even though, as we have seen, the need (and lack) of parallel and comparable corpora is still one of the biggest problems that research on multilingual language tools has to face.

There is a general tendency toward favoring the use of unsupervised machine learning techniques over non-annotated corpora for obvious economic reasons, but also because often there is no significant quality improvement in using annotated data (considering the time-consuming effort of semi-automatic and manual annotation). Furthermore the integration with linguistic resources in general is seen as an alternative way of adding linguistic awareness to purely statistical data processing. Some questions have been raised: Do really linguistics resources improve automating performance of translation tools and to what extent? It might be relevant here to draw a line between fully automated machine translation systems and CAT. With the exception of parallel and comparable corpora, other linguistic resources appear to be very expensive to build adding little or no relevant improvement to machine translation overall performance.

There are also problems relating the different performance quality of MT systems over languages with different morphological features. Morphological richness, like that of most Indo-European languages, still poses some problems in output quality. Are morphologically rich languages like romance languages as Italian still a step behind in MT?

The availability of textual data for languages other than English still poses some problems, even though the tendency in the digitization of all kinds of documentation is rapidly filling the gap and providing sufficient data for different training aims. Some of the more promising adventures in computational linguistics is taking place in the integration of different form of academic and commercial tools developed for very different purposes. The best example is WordNet, now widely used for a number of previously unpredictable applications. The open source nature of the Princeton project has been the key to its great success and this should be the best teaching, especially to academic research.

References

[1] De Mauro, T. (2009), "Basi di conoscenze e banche dati lessicali", in *Istituto della Enciclopedia Italiana. XXI secolo. Comunicare e rappresentare*, Istituto della Enciclopedia Italiana, Ed., Roma, pp. 253-308.

[2] Gandin, S. (2009), "Linguistica dei corpora e traduzione: definizioni, criteri di compilazione e implicazioni di ricerca dei corpora paralleli", *Annali della Facoltà di Lingue e Letterature Straniere dell'Università di Sassari*, vol. 5, pp. 133-152.

[3] Michel, J. B. *et al.* (2011), "Quantitative analysis of culture using millions of digitized books", *Science*, vol. 331, p. 176.

[4] Way, A. & Hassan, H. (2009), "Statistical Machine Translation: Trends & Challenges", presented at the 2nd International Conference on Arabic Language Resources & Tools.

[5] Chaudhuri, S. & Pino, J. (2009), "Recent Trends in Statistical Machine Translation."

[6] Koehn, P. (2005), "Europarl: A parallel corpus for statistical machine translation", *Proceedings of. MT Summit X*, vol. 5, pp. 79-86.

[7] Callison-Burch, C. *et al.* (2007), "(Meta-)evaluation of machine translation", *StatMT '07 Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 136-158.

[8] Callison-Burch, C. *et al.* (2008), "Further meta-evaluation of machine translation", *StatMT '08 Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 70-106.

[9] Mohler, M. & Mihalcea, R. (2008), "Babylon parallel text builder: Gathering parallel texts for low-density languages", *Evaluations and Language Resources Distribution Agency (ELDA).*

[10] Church, K. W. & Mercer, R. L. (1993), "Introduction to the special issue on computational linguistics using large corpora", *Computational Linguistics*, vol. 19, pp. 1-24.

[11] Resnik, P. (1998), "Parallel strands: A preliminary investigation into mining the web for bilingual text", *Machine Translation and the Information Soup*, pp. 72-82.

[12] Munteanu, D. S. & Marcu, D. (2005), "Improving machine translation performance by exploiting non-parallel corpora", *Computational Linguistics*, vol. 31, pp. 477-504.

[13] Talvensaari, T. *et al.* (2008), "Focused web crawling in the acquisition of comparable corpora", *Information Retrieval*, vol. 11, pp. 427-445.

[14] Hong, G. *et al.* (2010), "An empirical study on web mining of parallel data", *Proceedings of COLING'2010*, pp. 474-482.

[15] Almeida, J. J. & Simões, A. (2010), "Automatic Parallel Corpora and Bilingual Terminology extraction from Parallel WebSites", *3rd Workshop on Building and Using Comparable Corpora, lrec2010*, pp. 50-55.

[16] Resnik, P. (1999), "Mining the web for bilingual text", *ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* pp. 527-534.

[17] Resnik, P. & Smith, N. H. (2002), "The web as a parallel corpus", *Computational Linguistics*, vol. 29, pp. 349–380.

[18] Brandâo, A. M. *et al.* (2002), "Grabbing parallel corpora from the Web", *Procesamiento del Lenguaje Natural*, p. 13.

[19] Chen, J. *et al.* (2004), "Discovering parallel text from the World Wide Web", *ACSW Frontiers '04 Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation - Volume 32* pp. 157-161.

[20] Callison-Burch, C. & Osborne, M. (2003), "Bootstrapping parallel corpora", *HLT-NAACL-PARALLEL '03 Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond* pp. 44-49.

[21] Ma, X. & Liberman, M. (1999), "Bits: A method for bilingual text search over the web", *Machine translation summit VII*, pp. 538-542.

[22] Chen, J. & Nie, J. Y. (2000), "Parallel web text mining for cross-language IR", *Proc. of RIAO*, vol. 1, pp. 62-78.

[23] Baroni, M. *et al.* (2006), "WebBootCaT: instant domain-specific corpora to support human translators", *Proceedings of EAMT 2006*, pp. 247-252.

[24] Radu, I. *et al.* (2010), "On-line compilation of comparable corpora and their evaluation", *Proceedings of the Seventh International Conference on Formal Approaches to South Slavic and Balkan Langauges, Croatian Language Technologies Society; University of Zagreb Faculty of Humanities and Social Sciences*, pp. 29-33.

[25] Rapp, R. (1999), "Automatic identification of word translations from unrelated English and German corpora", *ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* pp. 519-526.

[26] Munteanu, D. S. & D. Marcu (2006), "Extracting parallel sub-sentential fragments from non-parallel corpora", *ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* pp. 81-88.

[27] Santos, D. (2002), "DISPARA, a system for distributing parallel corpora on the Web", *Advances in Natural Language Processing*, pp. 751-776.

[28] Potthast, M. *et al.* (2008), "A wikipedia-based multilingual retrieval model", *Lecture Notes in Computer Science*, pp. 522-530.

[29] Smith, J. R. *et al.* (2010), "Extracting parallel sentences from comparable corpora using document level alignment", *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pp. 403-411.

[30] Erdmann, M. *et al.* (2009), "Improving the extraction of bilingual terminology from Wikipedia", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 5, pp. 1-17.

[31] Tomas, J. *et al.* (2001), "Mining Wikipedia as a Parallel and Comparable Corpus", *9th International Conference on Intelligent Text Processing and Computational Linguistics*, vol. 1, p. 34.

[32] Yasuda, K. & Sumita, E. (2008), "Method for building sentence-aligned corpus from wikipedia", *AAAI 2008 Workshop (Wikipedia and Artificial Intelligence An Evolving Synergy* pp. 64-66.

[33] Erdmann, M. *et al.* (2008), "An Approach for Extracting Bilingual Terminology from Wikipedia", in Haritsa, J. *et al.* (eds), *Database Systems for Advanced Applications*. vol. 4947, Springer Berlin/Heidelberg, pp. 380-392.

[34] Nagata, M. *et al.* (2001), "Using the Web as a bilingual dictionary", *DMMT '01 Proceedings of the workshop on Data-driven methods in machine translation - Volume 14* pp. 1-8.

[35] Nerbonne, J. A. (1998), *Linguistic databases*. Stanford, CA, CSLI Publications.

[36] Resnik, P. *et al.* (1999), "The Bible as a parallel corpus: Annotating the 'Book of 2000 Tongues'", *Computers and the Humanities*, vol. 33, pp. 129-153.

[37] Resnik, P. & Melamed, I. D. (1997), "Semi-automatic acquisition of domain-specific translation lexicons", *ANLC '97 Proceedings of the fifth conference on Applied natural language processing* pp. 340-347.

[38] Steinberger, R. *et al.* (2006), "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages", *Arxiv preprint cs/0609058*.

[39] Koehn, P. *et al.* (2009), "462 machine translation systems for europe", *Proceedings of the twelfth Machine Translation Summit*, pp. 65-72.

[40] Tiedemann, J. & Nygaard, L. (2003), "OPUS — An open source parallel corpus", *NODALIDA 2003, the 14th Nordic Conference of Computational Linguistics*.

[41] Tiedemann, J. (2009), "News from OPUS — A Collection of Multilingual Parallel Corpora with Tools and Interfaces", *Recent advances in natural language processing V: selected papers from RANLP 2007*, p. 237.

[42] Zanettin, F. (2002), "CEXI: Designing an English-Italian Translational Corpus", *Language and Computers*, vol. 42, pp. 329-343.

[43] Bernardini, S. (2003), "Bi-directional Corpora and Translation: The CEXI Corpus", *TESOL Quarterly. Special Issue on Corpus Linguistics in TESOL*, pp. 528-537.

[44] Barbera, M. (2007), "NUNC-ES: New Tools for Corpus Linguistics in Spanish", *Cuadernos de Filologia Italiana*, vol. 14, pp. 13-32.

[45] Bernardini, S. *et al.* (2006), "A wacky introduction", *WaCky*, pp. 9-40.

[46] Baroni, M. & Kilgarriff, A. (2006), "Large linguistically-processed web corpora for multiple languages", *EACL '06 Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*, pp. 87-90.

[47] Baroni, M. *et al.* (2009), "The wacky wide web: A collection of very large linguistically processed web-crawled corpora", *Language Resources and Evaluation*, vol. 43, pp. 209-226.

[48] Wang, K. *et al.* (2010), "An overview of Microsoft Web N-gram corpus and applications", *HLT-DEMO '10 Proceedings of the NAACL HLT 2010 Demonstration Session*, pp. 45-48.

[49] Fellbaum, C. (1998),*WordNet : an electronic lexical database*. Cambridge, MA, MIT Press.

[50] Magnini, B. *et al.* (1994), "A Project for the Construction of an Italian Lexical Knowledge Base in the Framework of Wordnet", *SUN Microsystems*.

[51] Artale, A. *et al.* (1997), "WordNet for Italian and its use for lexical discrimination", *AI* IA 97: Advances in Artificial Intelligence*, pp. 346-356.

[52] Magnini, B. & Strapparava, C. (1997), "Costruzione di una base di conoscenza lessicale per l'italiano basata su WordNet", in *Linguaggio e Cognizione*, Carapezza, M. *et al.* (eds), Roma, Bulzoni.

[53] Roventini, A. *et al.* (2000), "ItalWordNet: a large semantic database for Italian", *LREC Proceedings 2000*, pp. 783-790.

[54] Roventini, A. *et al.* (2003), "ItalWordNet: building a large semantic database for the automatic treatment of Italian", *Computational Linguistics in Pisa*, *Special Issue, XVIII-XIX*, Pisa/Roma, IEPI, vol. 2, pp. 745-791.

[55] Bentivogli, L. *et al.* (2002), "MultiWordNet: developing an aligned multilingual database", *Proceedings of the First International Conference on Global WordNet.*

[56] Lenci, A. *et al.* (2010), "Building an Italian Framenet through Semi-automatic Corpus Analysis", *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pp. 19-21.

[57] Nirenburg, S. (1989), "Knowledge-based machine translation", *Machine Translation*, vol. 4, pp. 5-24.

[58] Knight, K. & Luk, S. K. (1994), "Building a large-scale knowledge base for machine translation", *AAAI '94 Proceedings of the twelfth national conference on Artificial intelligence (vol. 1)*, pp. 773-773.

[59] Nirenburg, S. *et al.* (1986), "On knowledge-based machine translation", *Proceedings of the 11th International Conference on Computational Linguistics COLING (1986)* pp. 627-632.

[60] Gangemi, A. *et al.* (2008), "The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in

WordNet", in *Proc. of On the Move to Meaningful Internet Systems OTM2003 Catania, Italy*, Springer Verlag, pp. 820-838.

[61] Oltramari, A. & Vetere, G. (2008), "Lexicon and Ontology Interplay in Senso Comune", in *Proceedings of OntoLex 2008 (Hosted by Sixth international conference on Language Resources and Evaluation)*, Marrakech (Morocco).

[62] Oltramari, A. & Vetere, G. (2008), "Acquiring Italian Linguistic Knowledge with Senso Comune", in *AI*IA 2008*.

[63] Oltramari, A. *et al.* (2010), "Senso comune", in *Proceedings of LREC 2010 7th International Conference on Language Resources and Evaluation*, May 17-23, Valletta, Malta.

[64] Chiari, I. *et al.* (in stampa), "Di cosa parliamo quando parliamo fondamentale?" in *Atti del Covegno della Società di Linguistica Italiana (Viterbo 27-29 settembre 2010)*, Roma, Bulzoni, in stampa.

[65] Roventini, A. & Ruimy, N. (2005), "Linking and harmonizing different lexical resources", *ItalWordnet and PAROLE-SIMPLE-CLIPS*, *GWC 2006 proceedings*.

[66] Roventini, A. (2006), "Linking verbal entries of different lexical resources", *LREC Proceedings*, *CDROM*, pp. 1710-1715.

[67] Navigli, R. (2005), "Semi-automatic Extension of Large-scale Linguistic Knowledge Bases", *Proceedings of 18th FLAIRS International Conference (FLAIRS)*, pp. 548-53.

[68] Navigli, R. *et al.* (2003), "Ontology learning and its application to automated terminology translation", *IEEE Intelligent Systems*, pp. 22-31.

[69] Navigli, R. & Velardi, P. (2007), "Glossextractor: A Web Application to Automatically Create a Domain Glossary", *AI* IA 2007: Artificial Intelligence and Human-Oriented Computing*, pp. 339-349.