Spammare senza pietà - Corpus based analysis of English, un-acclimatised verb loans in Italian and creation of a reference lexicon

Anna Fantini
Università degli studi di Pavia
anna.fantini01@universitadipavia.it

Abstract

English. We describe the lexical resource created to investigate the semantic changes of 90 English, un-acclimatised verb loans in Italian. Final results and interesting observations concerning the annotation task are discussed.

Italiano. *Descriviamo la risorsa lessicale creata per indagare in italiano il cambiamento semantico di 90 prestiti verbali inglesi non acclimatati. Illustriamo i risultati finali e le interessanti osservazioni emerse dall'esperimento di annotazione.*

1 Introduction

The case of language borrowing was investigated in depth by Gusmani (1983), who argues that a linguistic loan is an interference phenomenon, connected with contact and mutual influence of different languages. According to his study, the motivations behind the origin of a loan lie in the individual act of a speaker or of a group of speakers. The need to resort to a foreign alternative derives from the prestige held by the latter against an equivalent word in the mother tongue of the speaker (or from the absence all together of an alternative, as in our work: "*Se mi vede, Miki mi banna* (<to ban)" vs. *"*Se mi vede Miki mi bandisce*").

Facts show that language borrowing is particularly common among specialized languages, more so if they are linked to technical contexts.

This is extremely visible within the computer context. The main focus of this paper is the informal variety of Italian as used by communities of online video-gamers, computer experts and amateurs, forum users, etc.; a specialized language linked to technical context populated with partially integrated and un-acclimatised English verb loans.

These kinds of (mostly) lexical influences are so recent that their structure is hardly stable, and the process of integration – graphical, morphological, phonetic, and lexical – in the language is still in progress. For instance, they tend to retain the phonetic property of the original word, especially of the lexical root (to spawn > *spawnare* /spo'nare/).

The new word serves as an alternative – usually a hyponym – of an already existing term[1]. As for the concept of loan acclimatisation, the literature states that it involves the role of the new term in the target language. Therefore, Gusmani speaks of acclimatisation only with regard to the lexicon and its connection with speakers' usage: the more they familiarise with the loan, the more the latter gets acclimatised. It follows that – for very recent, scarcely integrated loans – the majority of speakers, as well as linguistic authorities, do not perceive the influence of English as an enrichment of the lexical heritage but mostly as a nuisance. If it is true that a number of reports describe the interference of English over Italian as an impoverishment, some attempts have also been made to study the less acclimatized loans themselves. It is thus of interest to examine why this kind of loan infiltrates the Italian language, how the speakers cope with the new word and what is the semantics of the loan in the target language. The aim of the present paper is to give a detailed account of how the meaning of a verb loan changes (and if it changes) in the target language and to offer a reliable source of lexical information in the form of an electronic lexicon built for the occasion. Section 2 details the method used to collect suitable data; section 3 illustrates the structure and functions of the lexicon; section 4 provides the results of our analysis as well as the annotation task performed with our data; section 5 discusses our findings and section 6 finally provides a conclusion.

2 Methods

In order to investigate the semantics of English un-acclimatized verb loans, we examined their occurrence in a monitor web-corpus created for

[1] E.g. *googlare* < to google as hyponym of *cercare.*

this purpose, following the guidelines and instructions of previous Corpus Linguistics works (Baroni and Kilgarriff, 2001; Lenci et al., 2012; McEnery, Xiao, Tono, 2006; Pomikálek, 2011; Pustejovsky and Stubbs, 2012). The corpus contains 6 transcriptions from a total of 194,07 minutes of audio material, collected with consensual but unaware recordings and then transcribed using the software Elan 4.9.6 (Wittenburg et al., 2006), plus 129 texts obtained through the Sketch Engine web-crawler, suitably set. We extracted a sum of 90 different verb lemmas (542 different word forms), for a total of 1327 occurrences. The annotation involves a POS level – limited to the sentence containing the loan – a loan-type level – describing three degrees of language integration[2] – a semantic type level[3] and a thematic role level[4]. The last two levels have been annotated using the tags proposed in Jezek and Nissim (2014) and Jezek and Vieu (2013) respectively.

Every text has been annotated using the Mae software (Stubbs 2011). An annotation task was conducted using a sample of the corpus (see section 4.2), its agreement result being only partially positive but interesting nonetheless from a linguistic point of view.

The next part of our research involved the analysis of the semantic patterns for each lemma[5], thus compiling one or more data-driven senses for every verb. The senses obtained were classified according to Verb Net's semantic class hierarchy. The assumptions underlying this investigation are grounded on Corpus Pattern Analysis (CPA) and Computational Lexicography (Hanks 2008; Hanks 2012; Jezek 2011).

Verb patterns have – in general – the following structure, where:

(1) *Spammare* 2b
Agent[PERSON] V_*spammare* (Theme[ARTEFACT | ABSTRACT]).

We have chosen all uppercase for the semantic type, and first letter uppercase for the thematic role, extended to every argument of the verb. Round brackets contain the possible optional arguments of the verb.

[2] Totally integrated, e.g. *spammare*; partially integrated (grafic), e.g. *trackare*; partially integrated (phonetics), e.g. *spawnare* /sp'nare/.

[3] E.g. Person, Artefact, Location, Abstract, etc.

[4] E.g. Agent, Patient, Goal, Source, Duration, ect.

[5] From Hanks, Pustejovsky 2005, a pattern is intended here as an argument structure with specifications of both the thematic roles and the semantic types of each argument positions.

3 The Lexicon

After extracting the semantic patterns for each lemma from the corpus, we stored the information in an electronic lexicon, built using the software Personal Lexicon 2.7.1, a language learning resource developed by Alexander Smith between 2007 and 2015. The software comes both in free and registered versions, the current lexicon has been compiled – and it will be consultable – using the free version.[6]

The lexicon is designed to give a precise account of every semantic feature and every meaning variation of the verb loans. As the reader will see observing Figure 1, each entry is characterized by the following elements (some of them pre-named in the software):

"Figure 1. The *spammare* 2b lexical entry"

- The entry citation form, with the number of the sense or of the sub-sense[7];
- The Pronunciation of the citation form;
- The Class (pre-named) as in the loan type which it belongs to (whether it is fully integrated or only partially integrated);
- The Root element, as in the lexical English root it comes from;
- The Theme (pre-named), as in the Verb Net class it was reduced to;
- The Definition box, containing the lexical definition and the verb pattern;
- The Related entries in the lexicon, all accessible through hyperlink;
- The Personal Examples, used to extract the pattern.

In figure 2 we show the Conjugations tab (pre-named) that includes all the syntactic complements of the verb and their semantic properties (thematic roles and semantic types).

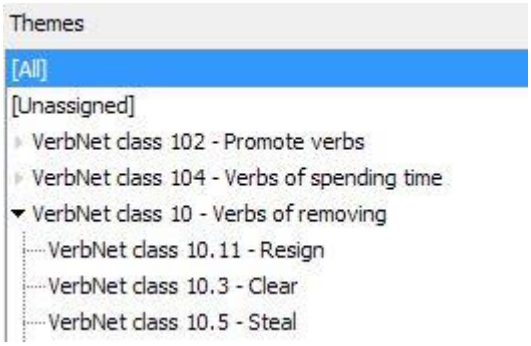
[6] The resource is not yet available for public consultation..

[7] Sub-senses corresponding to GRADIT's *accezioni*, the progressively numbered paragraphs a sense is pos-sibly divided into.

"Figure 2. Lexicon conjugation tab"

We listed the grammatical subject, the direct object, the indirect object and up to ten different indirect complements. Notice that the semantic type slot may specify more than one element, in which case we used the | separator.

In figure 3 we illustrate the Themes section of the lexicon, with a partial list of Verb Net classes and sub-classes used in the resource.



"Figure 3. Lexicon Theme section"

Clicking on each one produces the list of entries belonging to that particular class; this list appears in the third section of the lexicon, the Lexical Items column storing all the entries ordered alphabetically.

4 The Results

In this section we report the results of both the semantic analysis of the loans and the annotation task.

4.1 Quantitative considerations

The lexicon contains 157 senses for a total of 90 verbs. As shown in Table 1, the 157 senses have been classified into 3 groups according to three main criteria about the degree of semantic conservativeness of the loan:

1 The meaning remains the same as the original verb.

2 The meaning remains linked to the original one, but it diversifies to some degree.

3 The meaning changes to the point that it becomes a new meaning altogether

Group #	Type	Numbers
Group 1	Same sense	88
	New verb form	11
Group 2	Diversified sense	25
Group 3	New sense	26
	New v. and new sense	7
Senses	157	

"Table 1. Senses sorted according to their semantic behaviour"

Group 1 coincides with 63% of the total (99 senses out of 157), 11 senses have also new verb forms (*bishottare*, *autospottare*), 78 senses occur in 1 to 10 examples – they often have new verb forms (*riloggare*) or a very specific meaning (*rippare*)[8]. In this group there is the highest percentage (41%) of monosemic verbs.

Group 2 coincides with 15% of the total (25 senses out of 157), 19 senses occur only in 1 to 10 examples – they do not have a very specific meaning, but may be considered as hyponyms of Italian verbs (*farmare*2 of *sfruttare*)[9]. In this group there is the lowest percentage (16%) of monosemic verbs – their distribution proved not to be directly proportional to the senses' quantity.

Group 3 coincides with 25% of the total (33 senses out of 157), 24 senses occur only in 1 to 10 examples yet we also have the senses occurring in the highest number of examples (*droppare* 1[10] and 2[11] with 189 examples). In this group 27% of the verbs are monosemic.

4.2 The inter annotator agreement

The semantic annotation task was conducted following the methodology of Pustejovksy and Stubbs (2012); only a sample of 440 random occurrences was annotated by 9 groups of annotators, each constituted by 3 people. They were given guidelines explaining the method and the tagsets, and they were asked to separately annotate the semantic type and the thematic role of each verbal argument. We used Fleiss'*k* algorithm to calculate the agreement[12] (Artstein and Poesio, 2008), the values being interpreted according to Landis and Koch (1977). We already said that the results have been only partially positive, in particular – as for the thematic role – on-

[8] E.g. *"Non potendo accedere al CD-Rom non posso rippare niente"* .

[9] E.g. "*Ok, farmerò i campi di battaglia eterni*".

[10] E.g. "*Non molto tempo fa ho droppato i bracciali"*.

[11] E.g. "*Non è difficile droppare un computer privato"*.

[12] We choose Geertzen, J. (2012) online resource for agreement evaluation .

ly one group reached the 0.6 threshold considered acceptable with semantic annotation, the others showing moderate agreement and fair agreement (one group only). For the semantic type annotation, three groups reached the 0.6 value, four groups showed moderate agreement and two groups showed fair agreement. Nonetheless, we could make interesting linguistic considerations.

5 Discussion: the semantic behaviour of the loans

Let us consider the case of *spammare*[13]: all its three main senses are distributed among the three groups mentioned in Table 1, the semantic behaviour shows not only a certain degree of conservativeness, but also a great degree of diversification (just 18 occurrence out of 76 keep the original meaning) and thus of acclimatisation. Italian speakers apply a saving strategy: the monosemous loans are also the most conservative, while diversification often results in polysemous verbs – it seems that, once a semantic change starts, the speaker continues to use the loan until it reaches a definitive meaning, eventually becoming acclimatized. The cognitive effort behind this process is very high, but it also implies a certain linguistic confidence. Of course it is less arduous to produce a loan whose sense is strictly linked to the original verb's one, thus generating many monosemous loans. Nevertheless we wonder whether – aside from being economically convenient – is it also strategically and linguistically sensible to produce just monosemous loans instead of using semantically diversified ones. Is it sensible to keep numerous and specific loans, when there can be fewer and polysemous ones? Further investigations of English un-acclimatised verb loans may answer part of these questions.

5.1 Interesting observations about the annotation task

We feel that the only partially satisfying results may depend on the tricky lexical meaning of each loan. It is clearly easier to annotate the argument structure of a well-known verb like *potenziare*, rather than the one of the loan *overcloccare*[14] (*potenziare* and *overcloccare* being almost synonyms). The thematic role level is the most problematic, obtaining substantial agreement only in one case; the semantic type level on the other end is perceived as a less abstract, more transparent concept and the annotation is slightly better, with three groups over 0.6. What is really interesting is that the group which performed best with thematic roles is also the one which did worst with semantic types. Moreover, the groups which performed best with semantic types showed only fair to moderate agreement in thematic roles. We observe a – general and group wise – performance improvement with the semantic type level.

This is because assigning a thematic role requires a deeper reflection and some of the roles may be ambiguous (for example, Beneficiary and Goal). The creation and combination of more specific sub-types and sub-roles – targeted to this kind of verbs – could help resolve the ambiguity hindering agreement (for example, Person split into Authority and Subordinate)[15].

Furthermore the un-acclimatisation of the loans leads to somewhat different uses and different meanings among the speakers. This happens either between different communities, either between different speakers of the same community. Other significant observations emerged on the frequency of roles and types and on their co-occurrence: the most used roles are Agent (often erroneously) and Patient (often in the place of a more neutral Theme). The types most used are Person, Artefact and Abstract. The Agent-Person combination is the most frequent, even if the role is often wrongly assigned. Great uncertainty emerged in assigning the correct type to arguments whose referents are intangible informatics entities, e.g. nicknames, server, updates, etc. or characters of a game, e.g. boss, Pokémon, etc.

Last but not least, it was possible to already identify primitive verb classes, depending on the roles and the types assigned to verb arguments, i.e. verbs of change of state with a Patient role and possibly a Beneficiary role, or verbs of creation with a Result role and occasionally an Agent role.

---

[13] E.g. 1 "*Magari capiterà di spammare un oggetto"*
2 *"Iniziano a spammare pubblicità a tutti gli iscritti."*
3 *"Questa è un'abilità controversa [...] ma non va spammata"* .

[14] E.g. "*Prima di lanciare il tutto ho overcloccato la scheda video"*.

[15] From section 1 "*Se mi vede Miki mi banna*" (*bannare 1a)* Miki annotated as Agent-Authority and the personal pronoun as Beneficiary-Subordinate.

6 Conclusions

The peculiarities of each group of annotators lead to a thought-provoking analysis of the tagsets and the semantic notions themselves. The analysis of the semantic behaviour of the loans unveils deeper questions on the speaker's strategy: if it is easier to reuse a loan whose meaning already exists, why are there also loans with new or diversified meaning? Furthermore it came to our attention that yes, many loans are linked to the informatics/technical context (*rippare*, *sloggare*, etc.) but others can be considered as hyponyms of already existing Italian verbs, whose meaning is rather general (*cheattare* for *barare*, *whinare* for *lamentarsi*, etc.). It is possible to already discern synonymic and antonymic relations between the loans themselves: *craftare*, *farmare* 2, *spawnare* 1 or *sbuggare-buggare*. Finally, the most productive verb classes in our corpus are the verbs of change of state, the creation verbs and the verbs of killing.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555-596.

Joe Barcroft, Gretchen Sunderman and Norbert Schmitt. 2011. Lexis. In J. Simpson, ed. *The Routledge Handbook of Applied Linguistics*. Routledge Taylor & Francis Group. Oxon., pp. 597 – 610.

Marco Baroni. Italian tagset documentation. Available at:http://sslmit.unibo.it/~baroni/collocazioni/itwac.tag set.txt.

Marco Baroni and Adam Kilgarriff. 2001. *Large linguistically processed web corpora for multiple language.*

Raffaella Bombi. 2003. Anglicismi come banco di prova dell'interferenza linguistica. *Italiano e inglese a confronto, Firenze, Franco Cesati Editore.* 101-125.

Raffaella Bombi. 2005. *La linguistica del contatto: tipologie di anglicismi nell'italiano contemporaneo e riflessi metalinguistici.* (Vol. 11). Il calamo.

Tullio De Mauro. 1999–2007. *GRADIT - Grande Dizionario Italiano Dell'uso*. 6 vols. Torino: UTET.

Joseph L. Fleiss. 1971 . *Measuring nominal scale agreement among many raters*. Psychological Bulletin, 76:378-382.

Thierry Fontenelle. 2011. Lexicography. In J. Simpson, ed *The Routledge Handbook of Applied Linguistics*. Routledge, Taylor & Francis Group. Oxon. pp. 53 – 66.

Jeoren Geertzen. 2012. *Inter-Rater Agreement with multiple raters and variables*. Retrieved November 27, 2015, from https://nlp-ml.io/jg/software/ira/

Roberto Gusmani. 1983. *Saggi sull'interferenza linguistica (vol.1 e vol.2)*. Le lettere.

Patrick Hanks. 2012. How people use words to make meanings: Semantic types meet valencies. In A. Boulton & J. Thomas, eds. *Input, Process and Product: Developments in Teaching and Language Corpora*. Masaryk University Press.

Patrick Hanks. 2008. Lexical Patterns: from Hornby to Hunston and beyond. In E. Bernal & J. DeCesaris, eds.*Proceedings of Euralex 2008*. Barcellona, Universitat Pompeu Fabra.

Patrick Hanks and Elisabetta Jezek. 2010. What lexical sets tell us about conceptual categories. In *Lexis: E-journal in English lexicology,* 4: Corpus Linguistics and the Lexicon.

Patrick Hanks and James Pustejovsky. 2005. A Pattern Dictionary for Natural Language Processing. *Revue française de linguistique appliquée*, 10(2).

Martin Haspelmath. 2009. Lexical borrowing: concepts and issues. *Loanwords in the world's languages: a comparative handbook.*

Martin Haspelmath. 2008. Loanword typology: steps toward a systematic cross-linguistic study of lexical borrowability. *Aspects of language contact: new theoretical, methodological and empirical findings with special focus on Romancisation processes.* Mouton de Greuyter.

Elisabetta Jezek. 2011. *Lessico. Classi di parole, strutture, combinazioni*. Il Mulino, Bologna.

Elisabetta Jezek. 2010. Struttura argomentale dei verbi. In L. Renzi – G. Salvi (a cura di) *Grammatica dell'Italiano Antico*. Bologna: Il Mulino, 77-122.

Elisabetta Jezek and Malvina Nissim. 2014. *Linee guida per l'annotazione degli argomenti del verbo*

Elisabetta Jezek and Laure Vieu. 2013 . Lista di ruoli semantici per Senso Comune.

Adam Kilgarrif. 2001. Web as corpus. *Proceedings of corpus linguistics 2001.* Corpus Linguistics. Reading in a Widening Discipline.

Alan Kirkness. 2004. Lexicography. In A. Davies & C. Elder, eds. *The Handbook of Applied Linguistics*. Blackwell Handbooks in Linguistics. Blackwell Publishing, pp. 73 – 101.

Klaus Krippendorf. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications

Alessandro Lenci, Simonetta Montemagni and Vito Pirrelli. 2012. *Testo e computer. Elementi di linguistica computazionale*. Carocci editore.

Beth Levin. 1993. *English verb classes and alternations: a preliminary investigation*. University of Chicago Press.

Anthony McEnery, Richard Xiao and Yukio Tono. 2006. *Corpus based language studies. An advanced resource book*. Routledge, Taylor & Francis Group. Oxon. Routledge Applied Linguistics.

Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis. Brno, Repubblica Ceca: Masaryk University. Available at: http://is.muni.cz/th/45523/fi_d/phdthesis.pdf.

James Pustejovsky and Amber Stubbs. 2012. *Natural language annotation for machine learning*. O'Reilly.

Renata Savy. 2006. Specifiche per la trascrizione ortografica annotata dei testi raccolti. In *Progetto CLIPS. Corpora e Lessici dell'Italiano Parlato e Scritto.*

Amber Stubbs. 2011. *MAE and MAI: Lightweight Annotation and Adjudication Tools.*

Michael Stubbs. 2004. Language Corpora. In A. Davies & C. Elder, eds. *The Handbook of Applied Linguistics*. Blackwell Handbooks in Linguistics. Blackwell Publishing, pp. 125 – 152.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann and Han Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. In: *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.*