NeoTrack: semiautomatic neologism detection

Maarten Janssen

ILTEC, Lisboa

Abstract

NeoTrack is a web-based tool for the semiautomatic detection of neologisms in electronic corpora. NeoTrack was developed for the *Observatório de Néologia de Português* (ONP) to allow the daily observation of two major newspapers (*Diário de Notícias* and *Público*) for the occurrence of new words. This article describes the working of the NeoTrack application, its integration with the MorDebe database, and the criteria used in its application by the ONP.

1. Introduction

NeoTrack is a web-based tool for the semiautomatic detection of neologisms in electronic corpora. NeoTrack was developed for the *Observatório de Néologia de Português* (ONP) by the *Institúto de Linguística Teórica e Computacional* (ILTEC) to allow the daily observation of two major newspapers (*Diário de Notícias* and *Público*) for the occurrence of new words.

The disadvantage of computer-aided corpus-based neologism research is that computer tools are only capable of finding *formal* neologism (or in the case of NeoTrack - *orthographic* neologisms, see section 4.1). This because without semantic analysis it is impossible to tell the meaning of the words – and hence whether words are used in a new meaning. But the advantage of computer-aided research is not just that it saves a lot of time, but more importantly that it provides the means to establish (relatively) objective criteria about what a neologism is. Without the use of computers, it is virtually impossible to determine which words in a given text are really new – words may sound new without them actually being so, or sound familiar whereas they never occurred in any text before. This is why Rey (1975) in the pre-computer era said that to label a word neologistic is no more than the expression of a subjective sentiment.

With the use of computer-aided corpus research, it becomes possible to really establish which words are new by comparing the new text to the collection of all the text in a reference corpus. This makes it possible to find not just words that are completely newly created and also feel new, such as *pen-drive*, but also words from the potential lexicon that have recently become actualised. An example is the word *actor-chave*, which is a predictable word, not in the dictionary, which recently come into actual use (according to the criteria of ONP, described in section 4.2).

Because of its more objective character, computer-aided neologism research can be used for more than just updating dictionaries: the analysis of the neologisms obtained with NeoTrack gives an impression of the dynamics of the Portuguese language: which processes are most frequently used for the creation of new words, which languages are mostly used for new loanwords, which suffixes are most productive in new words, etc.

This article describes the NeoTrack application: its design and user-interface, an the way NeoTrack is integrated with the MorDebe database. Along side this article describe the criteria used in the application of NeoTrack by the ONP.

2. NeoTrack Design

NeoTrack is a light-weight tools for the observation of neologisms using a method of *exclusion based neologism candidate extraction*. This method says that a word in a text is possibly a neologism (a neologism candidate) when it does not appear in a list of previously known words, called the *exclusion list*. Neologism candidate extraction a semi-automatic process: the computer is used to generate a list of all possible neologisms – but it is up to a human user to decide whether these neologism candidates are indeed neologisms or false candidates. Although this latter step could in principle be made automatic, fully automatic neologism extraction is more commonly fully stastics-based without the intervention of a neologism candidate list.

The way neologism candidate extraction is implemented in NeoTrack is illustrated in figure (1): to extract all the neologism candidates from a given text (*corpus file*), the system first creates a list of all the unique words occurring in that text (*corpus words*) by tokenising the cleaned-up version of the corpus file (*corpus text*). This list is then compared with a list of known words (*exclusion list*) to render a list of all the words the system does not recognise: the *neologism candidates*.

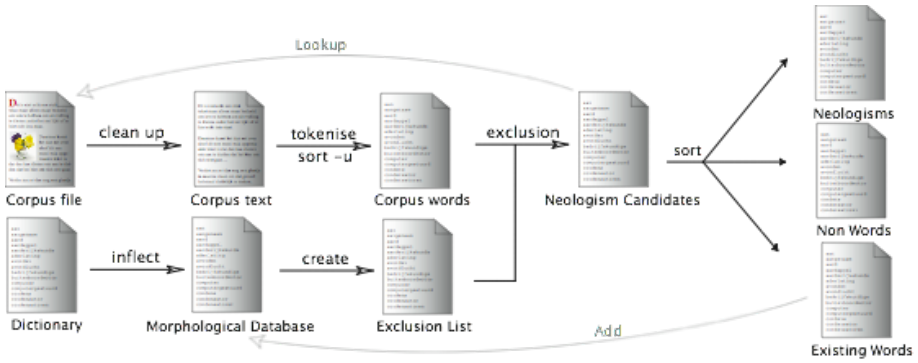


Figure 1: Neotrack flow-chart

The exclusion list in NeoTrack is created from a morphological database called MorDebe, which itself is derived from lexicographic resources. The MorDebe datbase in turn is created partially from lexicographic sources (see section 2.1). The exclusion list in NeoTrack does not contain only the citation forms of all the words, but also all the inflectional forms.

NeoTrack not only extracts the lists of neologism candidates, but features a user-frienly interface to split the neologisms from the *false candidates*. False candidates are those unknown words that are not neologistic, either because they are existing words that were missing from the exclusion list, or because they are strings that should not be counted as words: proper names, typographic errors, etc. The user interface (see section 3) is fully web-based an can be accessed via any Internet browser. The use of a server-based system allows the linguists to work from any computer they want – even allowing neologism observation from an Internet café. This is not merely a convenience, but it allows researchers of different institutes, and even of different countries speaking the same language, to cooperate in a single project, working with the same neologism database.

2.1 Integration with MorDebe

NeoTrack is integrated with a morphological database called *MorDebe* (Janssen 2005a; Janssen 2005b). MorDebe is a large-scale lexical resource which contains a large amount of correct Portuguese words, including all their inflected forms. MorDebe is an online service that works as an orthographic guide, a verb dictionary, and an inverse dictionary – with a rich set of search options.The design of the database is language-independent, but only data for Portuguese are available for the moment. The aim of MorDebe is not to provide as many words as possible, but to provide a lexicographically controlled lexicon with manual verification at every point. The database started with a semi-automatic inflection of the lemmas of the *Porto Editora* dictionary, but has since been updated with words from various sources including the CETEMPublico corpus, the Academia and Houaiss dictionaries and the NeoTrack research. At this moment, the database contains well over 125.000 lexical entries for Portuguese, with an emphasis on the European variant of Portuguese.

MorDebe is not just used in NeoTrack, but was even originally conceived for the purpose of the ONP neologism observatory with NeoTrack - and the two system are fully integrated. On the one hand, the exclusion list used in NeoTrack is created on-the-fly from the MorDebe database: just before extracting the exclusion list from the corpus words to create the neologism candidates, the exclusion list is (optionally) updated with all the word-forms in the MorDebe database, to also exclude the most recently added words. In this way, the observation of neologism speeds up with the growth of MorDebe because the number of neologism candidates will diminish.

On the other hand, NeoTrack is used as one of the methods to keep MorDebe up-to-date: when a linguist in the use of NeoTrack encounters a word that is not a neologism, but an existing correct word that was somehow missing from MorDebe, that word

can be directly added to MorDebe from the interface of NeoTrack (after corpus verification). Also, the MorDebe database is periodically updated with all those words from the neologism data base created by NeoTrack that turn out not to be occasionalisms.

3. User Interface

The NeoTrack user interface is divided into three major parts: the management of source files, the neologism candidate sorting, and the neologism database itself. This section gives a brief overview of the design of these three parts.

3.1. File Management

NeoTrack is a web-based system, which means that all files to be processed need to be uploaded to the server first. Processing a corpus file therefore happens in two steps: in a first step, the file is uploaded from the local computer to the server, and stored in the list of files to be processed. And in a second step, the corpus file is analysed, and the list of neologism candidates is extracted. The final step of this process leads to a list of ni progress, as shown in figure (2). Each file is shown with its source – and the amount of neologism candidates encountered in the text – with an indication of how many candidates have yet to be processed. With each file, as with all other data in the system, NeoTrack keeps track of which user added the file, and when he/she did so.

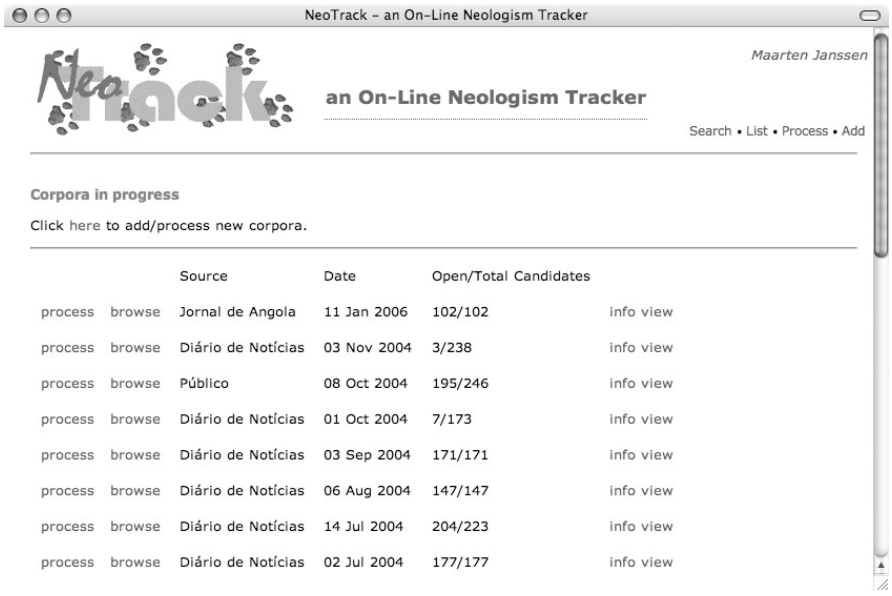


Figure 2: File Management

For each corpus in progress, it is possible to start/continue the process of neologism candidate sorting (see next section), or view the list of all the candidates of that corpus together with their status: *open*, or an indication of the action performed on the candidate. To get more information about the corpora, it is also possible to view the frequency distribution list of all token words in the different corpora, or view the original HTML file.

In the design of NeoTrack, the comparison between the input text and the exclusion list is done only once. This means that if a word is added to the exclusion list after the candidate list has been created will not affect existing candidates list. Therefore, it is recommendable to keep files in the unanalysed until it is actually being treated.

3.2. Neologism candidate sorting

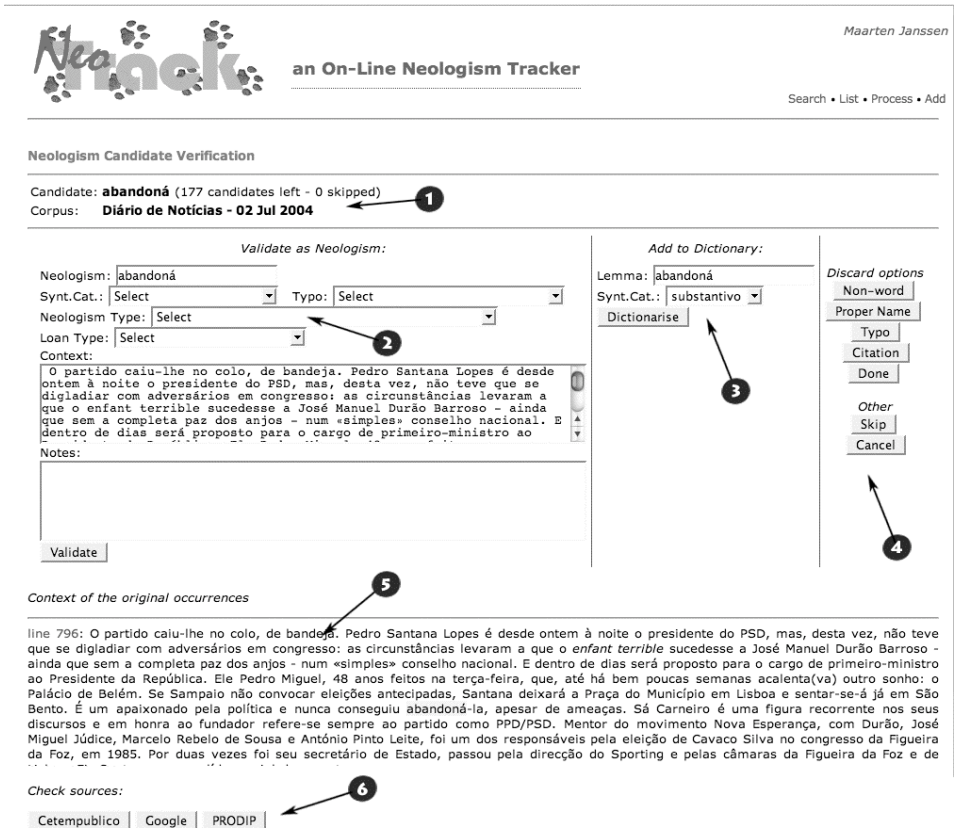


Figure 3: Candidate sorting

The main window for the manual sorting of neologisms is shown in figure (3), with circles added to indicate the main components. Every neologism candidate carries with it the spelling of the neologistic form, as well as the source in which it was encountered. This information is shown under (1). Next to that is an indication of the number of candidates in that source that have not been processed yet. To decide whether a candidate is a neologism, the original context is shown under (5) – where clicking on the line number will display the original HTML file to see the entire context. If the candidate appears more than once, multiple context lines are displayed.

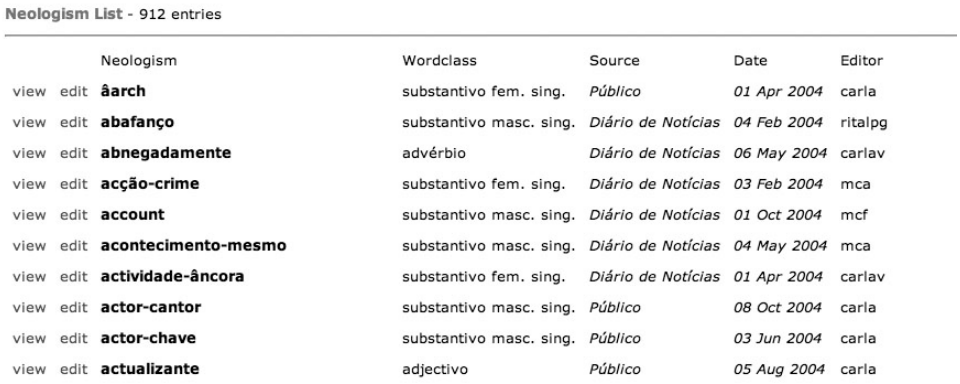
The main purpose of the sorting window is to allow the user to decide whether the neologism candidate is indeed a neologism or not. When the candidate is a neologism, the relevant data about that neologism can be entered under (2) – the citation form, syntactic category, its typography, and neologism type. The context in which the neologism occurred is automatically selected – but can be edited when the context is longer or shorter than desired. When the same candidate appears various times in the same source, the context of the first occurrence is selected. When validated as a neologism under (2) the candidate will be put in the neologism database, with all the associated data.

When the candidate is not a neologism but a false candidate, it will not be stored in the neologism database, and can be discarded. There are several reasons for discarding a candidate as a neologism – which are shown under (4): the candidate can be a typographic error, or a proper name. It can be a part of a foreign-language quotation, or it can be something which is not a word – such as an e-mail address, a code, etc. All the buttons under 4 do the same – but the motivation for rejecting a neologism is kept on file, to be able to use that information later, for instance to select all proper names. It is also possible to postpone a specific candidate until later in case there is some doubt about it.

Finally, the candidate can also be non-neologistic because it is an existing correct word, but just one that was not yet on the exclusion list. In that case, the word is not only removed from the candidate list, but added to the exclusion list so that it will not show up as a neologism candidate again. Since MorDebe is used for the creation of the exclusion list in NeoTrack, the word can be directly added to MorDebe under (3) – by indicating citation form and word class. Clicking on 'Add' will open the MorDebe administration page, where not just the particular form occurring in the source, but the entire inflectional paradigm of the word will be added to the MorDebe database. To decide whether a word is new or old, it is necessary to consult reference corpora. Therefore, under (6) are some quick links to look up the candidate in some on-line corpora.

3.3. Neologism database

In the neologism database section, it is possible to view and edit all the neologism already stored in the neologism database. An example from the ONP neologism list is shown in figure (4).



Neologism List - 912 entries

		Neologism	Wordclass	Source	Date	Editor
view	edit	**âarch**	substantivo fem. sing.	*Público*	*01 Apr 2004*	carla
view	edit	**abafanço**	substantivo masc. sing.	*Diário de Notícias*	*04 Feb 2004*	ritalpg
view	edit	**abnegadamente**	advérbio	*Diário de Notícias*	*06 May 2004*	carlav
view	edit	**acção-crime**	substantivo fem. sing.	*Diário de Notícias*	*03 Feb 2004*	mca
view	edit	**account**	substantivo masc. sing.	*Diário de Notícias*	*01 Oct 2004*	mcf
view	edit	**acontecimento-mesmo**	substantivo masc. sing.	*Diário de Notícias*	*04 May 2004*	mca
view	edit	**actividade-âncora**	substantivo fem. sing.	*Diário de Notícias*	*01 Apr 2004*	carlav
view	edit	**actor-cantor**	substantivo masc. sing.	*Público*	*08 Oct 2004*	carla
view	edit	**actor-chave**	substantivo masc. sing.	*Público*	*03 Jun 2004*	carla
view	edit	**actualizante**	adjectivo	*Público*	*05 Aug 2004*	carla

Figure 4: Neologism database listing

Each candidate is shown with the source it was encountered in, and the person who treated the neologism. By clicking on *view* it is possible to see all the data associated with an individual neologism. It is also possible to search the neologism database on all the various fields, or edit erroneous items in the neologism database.

4. Identifying Neologisms

An important aspect of the detection of neologisms is a proper specification of which words do count as neologisms. Although there are various ways of defining what a neologism is, the definition used by the ONP is called the *extended lexicographic diachronic criterion* (Janssen, *unpublished*). This criterion is a hybrid of the traditional *lexicographic criterion* and the *corpus-based criterion* (Cabré, 1992).

On the one hand the hybrid criterion is dictionary based in the sense that it uses dictionaries for its exclusion list: any word appearing in the dictionary is not a neologism. The dictionaries used for this purpose by the ONP are the *Porto Editora, Houaiss*, and *Academia* dictionaries (see 4.2). Rather than using the dictionaries directly, the system is based on a morphological database explicitly listing all inflected forms of all the lemmas, information often left implicit in dictionaries.

But on the other hand, appearing in the dictionary is only taken as a sufficient condition for being a correct word. Dictionaries leave out many words because of spatial limitations, especially semantically transparent words such as compounds and regular derivations. Since those words are not sensibly considered new, the hybrid criterion depends on the use of corpora to determine whether a word is absent from the dictionary because it is too new, or only because of the lexicographer's choices to leave it out. Therefore, the hybrid criterion relies on the lexicographic method rather than on any given lexicographic product.

4.1 Ortographic Neologisms and Neologistic Occurrences

By the design of the system, NeoTrack is not capable of detecting all forms of neologisms, but only what are called *orthographic neologisms* (Janssen, *unpublished*). Since the system depends only on the orthography to create the list of neologism candidates, it cannot detect semantic or pragmatic neologisms, but only formal neologisms. But even stronger - it is only possible to new strings, since not even word class is taken into account. This would in principle mean that even zero derivations would not be detectable by the system. But as a rule, different word-classes as a rule inflect differently, and in NeoTrack it is possible to encounter a neologism by any of its inflected forms, because NeoTrack does not perform any lemmatisation. Therefore, orthographic neologism are slightly less restrictive than pure string-based neologisms.

The neologisms stored in the neologism database do not have the restriction of permanency: any odd occurrence of a correct word will count as a neologism. For that reason, the items in the neologism database are more correctly referred to as *neologistic occurrences*, since other than the notion of neologism in dictionaries, these are simple occurrences, without any judgement on whether the word is likely to become an established word or is a clear occasionalisms.

4.2. ONP Criteria

Although the use of NeoTrack makes the detection of neologisms much more objective than the manual collection of terms, there are still several factors in the process that are arbitrary – when to count a word as established, which forms to count under the inflectional paradigm, etc. For all these free variables, standards were set within the ONP to reach a higher level of objectivity (Correia *et al.* 2004). These standards were not chosen at random, but reflect the standards of the internal consortium NeoRom – a network of neologism observatories for all Romance languages, using similar criteria, with the goal to reach comparable neologism database for the different languages. This is turn with the objective to allow a comparison of the change of the different language, possibly leading to a common standard for the incorporation of loanwords. This section described the most prominent arbitrary standards used by the ONP.

The first and maybe most important standard is the periord during which a word counts as a neologism: although the words *DVD* and *pen-drive* are both new, the second is much newer than the first – newness is a gradual notion. As a standard – a word in the ONP counts as a neologism if it first appear less than 3 years ago. On the one hand, this means that no texts in the reference corpus may be less than 3 years old – and that no text under consideration should be more than 3 years old.

The second issue is which words are considered established Portuguese words. Within the ONP established words include all the words occurring in one of the main (European) Portuguese dictionaries: Houaiss, Porto Editora and Academia. But also all those words that occur correctly at least 10 times in the collection of corpora used as a

reference corpus. At the moment, these include the CETEMPublico corpus, the REDIP corpus, the CLUL corpora, and the Linguateca AC/DC corpora.

A third issue is the notion of an inflectional paradigm – this because from the perspective of the dictionary, the adverb *arrevezadamente* is considered to be implied by the dictionaries, because the adjective *arrevezado* is in the dictionary. But from the perspective of langugage change, these are exactly the most productive mechanisms for the creation of neologisms. Therefore, all derivations including *inherent inflections* (Booij, 1995; Janssen, 2005) are considered potential neologisms. An exception is the augmentative of adjectives, which is without debate derivational, but still considered fully productive in Portuguese. This last rule is only overruled for non-gradable adjectives.

5. Conclusion

The NeoTrack application provides a user-friendly way extract neologisms from corpora. The tools provides everything necessary from beginning to end – the only thing left for the user to do is to collect the HTML sources of the corpus he wants to analyse, feed them to the system, and judge all the neologism candidates manually.

The basic design of the NeoTrack system is independent of language and neologism criteria. But when used in combination with the MorDebe database the system automatically uses the extended lexicographic diachronic criterion, which defines words as neologistic only when they (1) do not appear in the dictionary (MorDebe), (2) are correct words, and (3) do not occur above a threshold frequency in the reference corpus.

The use of the NeoTrack system for the observation of neologisms leads to a well-founded database of all new words of the language, which in turn gives an insight into the productivity of the language in terms of most productive suffixes, the most common loanword sources, etc. The neologism database produced by NeoTrack contains neologistic occurrences – which might be occasionalisms. To derive a neologism dictionary from the database, only those items that appear above a threshold frequency in the neologism database itself. Because of the integration with MorDebe, the observation of neologisms not only leads to a neologism database, but also to an enriched morphological database.

References

Booij, Geert (1995) Inherent versus contextual inflection and the split morphology hypothesis. *In:* Booij & van Marle (eds.) *Yearbook of Morphology 1995*. Dordrecht: Kluwer.

Cabré, M. Teresa (1992) *La Terminología: Teoría, metodología, aplicaciones*. Barcelona: Editorial Antártida.

Correia, Margarita; Mafalda Antunes, Ana Mineiro, Maria Doria & Teresa Cabré (2004) O Observatório de Neologia do Português Europeu – ONPE: criação e apresentação. *Actas do XIX Congresso da Associação Portuguesa de Linguística (APL)*. Lisboa, Portugal.

Janssen, Maarten (2005a) Between Inflection and Derivation: Paradigmatic Lexical Functions in Morphological Databases". *Proceedings of MTT-2005*, Moscow, Russia.

Janssen, Maarten (2005b) Lexical vs. Dictionary Databases.: design choices of the MorDebe system. *East-West Encounter: second international conference on Meaning-Text Theory*. Moscow, Russia.

Janssen, Maarten (unpublished) *Orthographic Neologisms: selection criteria and semi-automatic detection*. URL: http://maarten.janssenweb.net/publications

Rey, Alain (1975) The Concept of Neologism and the Evolution of Terminologies in Individual Languages. *In:* Sager (ed.) *Essays on Terminology*. Amsterdam: John Benjamins publishing. Translation of: *L'aménagement de la Néologie*. Office de la language française du Québec. 1975, p 9-28.