Šime Demo

Mining macaronics

1. Introduction

1.1. Macaronic Latin

There is no universally accepted definition of the macaronic style. The term was first employed by Michele “Tifi” degli Odasi in the late 15th century to describe a curious mixture of Latin and a northern Italian dialect in which Latin endings were attached to vernacular stems with excessive frequency, verse was hexametrical, and themes were usually burlesque, lampooning classical epics.1 It was sometimes also used to describe the language of all multilingual, humorous, or unconventional works (such as tautograms). More recently, with the development of medieval studies in the 20th century, medievalists started to use the term to designate linguistically mixed texts from that period, regardless of the way in which languages combined.2 Recently, there have been attempts to comprehensively embrace both traditions.3

In the present paper I use the original meaning of the term, which designates the language of the literary genre that arose from the humanist culture, quickly spread throughout Europe, lasted for several centuries, and included about a dozen language pairs. This implies that the style was employed intentionally, that it is characterised by densely distributed lexical hybrids with vernacular stems and Latin endings, that the dominant language is Latin, that the content is usually humorous, and that the macaronic poetry itself is an offshoot of the humanist culture (in that, for example, it applies classical metrical form or regularly references ancient classical works). Although such works form a clearly defined tradition, admissible outliers exist in macaronic prose (as found in the introductions to macaronic

---

3 E.g., Boehme (2011), in which the author covers a span of 1000 years of what she considers English macaronic texts. Concerning the problem of the definition, see Demo (2014).
poems in particular) as well as in non-macaronic works that nonetheless contain macaronic parts (such as Rabelais’s *Gargantua and Pantagruel* and Moliere’s *Imaginary Invalid*). Occasionally, integral vernacular words are likewise used.

One example of the Neo-Latin macaronic style, found in verses written by the greatest macaronic poet, Italian Teofilo Folengo (1491–1544), is given in (1):

(1) **An poterit passare maris mea gondola scoios,**
    quam recomandatam non vester aiuttus habebit? (Cocaius [Folengo] [1552] 1911: v. I.7–8)
    ‘Will my gondola be able to pass the rocks of the sea, if it will not be entrusted to your assistance?’

The Neo-Latin macaronic style has no fixed vocabulary, even within a single language pair; nonetheless it is predictable because of the unequivocal rules of its formation. In this sense, it transcends mere code-switching and instead places itself on the cusp of a completely novel idiom, i.e., an original mixed language. If observed more straightforwardly as characterised by code-switching, however, Neo-Latin macaronics formally comes close to what Muysken (2000: 122–153) terms “congruent lexicalisation”.

### 1.2. Status quaestionis and research problems

Although Neo-Latin macaronic poetry is a case—and an eccentric one, at that—of code-switching at work, it has seldom been analysed with regard to its structural properties in general and to switching in particular. Most of the research focuses on philological and literary problems or historical context, complemented by sporadic analyses of the sociolinguistic framework.\(^4\) Those few studies that undertake close linguistic analysis have focused on just one language pair at a time.\(^5\) Additionally, most of the authors did not have the advantage of using computer technology in their research.

Neo-Latin, the dominant language in macaronics, is a very specific linguistic form. It has many features of an artificial (i.e., not completely natural) language, because educational and literary conventions have always demanded the use of set phrases from classical Roman

---

\(^4\) The body of scholarship in the field has grown considerably; a good list of selected references is found in Sacré (2007).

antiquity. Macaronic Neo-Latin is even more artificial, because its own rules must be superimposed upon the rules of Neo-Latin and consciously followed in order to achieve the desired effect.

All Neo-Latin macaronic works, regardless of the Embedded Language (hence: EL), have a common linguistic basis and a unique overall makeup. Because they constitute a finite corpus, it is possible to compare them, to identify their rules of composition, and to observe the flexibility of as well as the occasional exceptions to those rules. The present study focuses primarily on building a foundation for the digitally supported, integrated, and comparative linguistic study of the entire Neo-Latin macaronic tradition. Since digital technology facilitates the automatic processing of large amounts of data, its application may represent a considerable methodological improvement on existing studies in the field. In addition, it can assist in the identification of marginal, theoretically problematic, and ambiguous macaronic phenomena.

A two-stage approach is taken here. Firstly, a sample corpus is built and prepared for analysis, i.e., annotated, exploring along the way the extent to which computers can automate the relevant tasks. In the present study the annotation is made on the level of the word, which is but one possible option.

Secondly, two sample analyses on the corpus are carried out. This involved an observation of how various word types (Latin, hybrid, EL) cluster, by means of calculating the sizes of the continuous monolingual text segments. Special attention is given to possible similarities between texts with closely related ELs. Similarly, a test is run to see whether verse beginnings play an important role in achieving the macaronic effect.

1.3. The corpus and the methodology
There are at least 130 Neo-Latin macaronic authors that appear in widely available literature. The present corpus consists of extracts from 60 macaronic poems. It includes 11 different language pairs, with Latin alone represented throughout. The ELs are all Indo-European, equally representing three major European linguistic subfamilies: Germanic (Dutch: 18, Latin gives most of the grammatical morphemes and dominates quantitatively; therefore, it can be conveniently termed the Frame Language (hence: FL) and its partner language in macarons, the Embedded Language (hence EL), although these do not behave in exactly the same way as naturally occurring FLs and ELs as described in Myers-Scotton (1993).

7 Counting individual works may be misleading: for example, a single short poem cannot be equaled to a long epic or a series of epigrams, although each of them can be represented by one bibliographical entry.
8 Figures represent the numbers of poems included in the corpus.
English: 7, German: 7), Romance (French: 5, Italian: 13, Portuguese: 5, Provençal/Occitan: 5, Spanish: 5), and Slavic (Croatian: 1, Czech: 1, Polish: 10). Within individual language pairs, the poems were selected randomly.

Typical Latin macaronic works are scarcely recognized outside the listed linguistic combinations. In terms of authorship time span, the earliest writing is from c. 1484 and the most recent is from 1969, but the majority date from the 16th to 18th centuries.

Our corpus has 3,274 verses in total, with chunks containing an average of 54.57 verses. The total word count is 20,603, which averages out to 343.42 words per poem. The topics are various: political, religious, erotic, intellectual end so on. The narrative framework ranges from playful occasional poetry to long epic-like parodies. The authors are also very diverse—we encounter nationally important poets such as Teofilo Folengo, Jan Kochanowski, Remy Belleau, and Tomás de Iriarte, but also less known names such as Bartolomeo Bolla, and João da Silva Rebello, or even people who made a name in other scholarly fields such as historian and philologist Antonio Muratori, theologian Alexander Geddses, or biologist Edward Forbes, as well as several anonymous works.

Hunston (2009: 160) lists three methodological requirements for a corpus: representativeness, balance, and size. The corpus employed herein is representative (all periods as well as most of the existing language combinations and major language families are covered); it does not contain, e.g., Basque, Finnish, Estonian, or Hungarian, however, because Neo-Latin macaronic works in those languages, if they exist, have not been widely accessible to the Neo-Latin research community. Meanwhile, as some languages (namely Croatian, Czech, and Dutch) have relatively modest macaronic traditions, their parts of the corpus are considerably smaller and the sizes of their chunks are relatively large, while the extracts from other texts are of the approximately same size. The number of poems selected from most language pairs is roughly equal, with more space given to larger traditions, especially to the Italian, which is by far the most comprehensive one. The results obtain by the statistical analysis have been normalised when necessary. Furthermore, although as mentioned there do exist some instances of macaronic prose, only macaronic poetry is included in the present analysis, because it presents the basic form of macaronic language and ensures stylistic consistency across the data. Any comparison of macaronic prose and poetry should take into consideration the differences that generally exist between the language of (Neo-)Latin poetry and its prose counterpart in terms of word choice and order.
The size of the corpus is an especially troublesome matter. The texts are not very numerous because they are mostly unavailable in digital form, and this is for several reasons. Firstly, the texts predate digital publication; secondly, Neo-Latin works are not nearly as often digitised as ancient ones or those written in major modern languages; finally, even within the Neo-Latin tradition, macaronic works are relatively marginal and have thus been only sparingly digitised.

Nevertheless, even a corpus of this size is sufficient to demonstrate the current methodology, as the conclusions of our analysis are not based on low-frequency items. As the case is with any corpus, the larger it becomes, the wider its applicability gets.

Although the corpus is comprised of texts from the past, given the very limited range of diachronic linguistic change in Neo-Latin, the corpus can be said to be historical only regarding the ELs. Thus it can be viewed as a group of historical texts that are comparable on a certain basis.9

If a computer is to be used for the processing of text, all characters of the text have to be encoded according to a character encoding scheme. The most basic scheme, the so-called ASCII, can encode all characters of the English (and Latin) alphabet, but is unable to deal with many characters that appear in, for example, French (ç), German (ä, ü, ö), Polish (ł, ś), or Croatian (ć, č, š). A more recent scheme, UTF-8, which can handle more than a million characters, is used in the present corpus. If, e.g., Russian, Greek, or Georgian are to be added in the future, UTF-8 can handle their respective scripts, too.

For text tagging XML encoding is used, as it is a widely accepted standard.10 Shorthand titles were added in front of each poem. The lines containing the titles were tagged with <title>Some title</title>, which serves as a unique delimiter of corpus poems.

Annotated in this way, the corpus becomes a quantitative and qualitative data mine, and as such is able to reveal some high-level properties of macaronic language. The

9 See Aijmer (2009: 276).
10 XML (Extended Markup Language) is a standard markup language that enables both humans and computers read a tagged text. All metadata are placed in a tag, called element, that encloses the part of the text it refers to. Elements can contain further details, which are stored in the so-called attributes. For example, a description of the Italian word verso can be encoded like this: <word lang="ita", meaning="verse", pos="noun", gender="masc", number="sing">verso</word>, where the element <word>, giving information about the word verso, contains attribute lang with the value "ita", attribute meaning with the value "verse", and so on. The main advantages of the XML are its flexibility and platform-independence.
programming language used in text encoding, Perl, also proves convenient in analysis.\textsuperscript{11} It supports very complex pattern searching (via the so-called regular expressions), which is used in retrieving data.

2. Building the corpus

2.1. Binary tagging

As already mentioned, the word is the operative level of analysis in the present work. The TEI-recommended\textsuperscript{12} \texttt{<w>}</w> element for each word is used, to which two attributes are then assigned: \texttt{lang} for the languages involved, and \texttt{type} for the classes of the word tailored to the corpus. Words are identified with the help of regular expressions, and after some minor manual cleanup, every single word in the corpus is tagged with the element of the following type:

\[
\texttt{<w lang="xx" type="xx">some_word</w>}
\]

where "xx" is added as a placeholder for the attribute values, in order to anticipate the final structure of the element.

The rest of the corpus preparation consists mostly of filling in and adjusting the attribute values within the word elements (see Figure 1 for the overview of the process). The first step here is distinguishing between the two major word types that appear in macaronic poetry: Latin words and non-Latin (macaronic) words. Of course, we want, whenever possible, a computer program do the tedious work of telling them apart. This is a matter of a straightforward comparison between a list of Latin words and the words in the corpus texts. As Latin is a highly inflected language, and it is desirable to account for every possible word, a list of not only lemmas but of all word forms is needed. In order to obtain this one can use a lemmatiser, which is a program that checks an input against a database of word forms and, if the word from the text is found, returns its lemma and possibly other information.\textsuperscript{13} Lemmatisation per se is unnecessary here, as the objective is simply identifying words as Latin, but since lemmatisers are connected with databases comprising Latin word forms, they provide the necessary information.

\textsuperscript{11} See Hammond (2003) and Bilisolly (2008), books specialising in linguistic research with Perl, which offer some good ideas for the analysis. I would like to thank my colleague Jan Šipoš for having drawn my attention to the usefulness of Perl in linguistic analysis and to the existence of the Morpheus lemmatiser (see below).

\textsuperscript{12} TEI (The Text Encoding Initiative), as its web site states, “is a consortium which collectively develops and maintains a standard for the representation of texts in digital form” (see TEI: Text Encoding Initiative 2013). Its standards are described in TEI Guidelines and use XML as the markup language.

\textsuperscript{13} There are several lemmatisers for Latin available on the Internet, e.g., Collatinus (Ouvrard & Verkerk 2014), or Morpheus, made by the the Perseus Project (Crane 2015); the latter is used in the present study.
Another method would be building from scratch a comprehensive form database, which can be generated from a list of dictionary headwords with unambiguous information about the class of their inflection. Such lists are available on the Internet, and it is possible to produce the remaining forms. If the database and the text to be processed are stored on the same computer, transmission of data across the Internet is unnecessary, and the program executes more quickly.

Of course, the output in such a case would depend on the varieties of Latin covered with the form database. As will be demonstrated further down, ready-made databases of Latin often contain forms that are superfluous to a given search or, conversely, lack forms that are needed. One of the advantages of using a self-built database is that it is possible to predetermine which words and forms are included in a given search. As this process could be extremely time-consuming, an alternative would be to generate the database from an appropriate collection of the existing Latin texts.
The lemmatiser gives each word a binary value: it is either Latin or non-Latin. If the word is [+Latin], "la" replaces the dummy lang attribute value;\textsuperscript{14} if it is [-Latin], "xx" is left unchanged as an EL code placeholder (more on this below). Furthermore, since the program detects pure Latin words, the attribute value "p" (< lat. purum) is set as their type attribute value. Thus, the Latin word ergo ‘therefore’ now looks like this: \textless w lang="la" type="p">ergo\textgreater , and the hybrid word dansas ‘dances’ (n.), like this: \textless w lang="xx" type="xx">dansas\textgreater . As XML code is relatively hard to read, the texts can be transformed into HTML and, thanks to the newly added tags, automatically formatted according to the word types. In the verses given in (2) non-Latin words are boldfaced:

\begin{align*}
(2) \quad & \text{Qui volet ergo bassas apprendere } \textbf{dansas} \\
& \text{Et cito } \textbf{dansandi mestrus} \text{ in arte fore,} \\
& \text{Hunc bene de testa se } \textbf{forcet} \text{ discere librum,} \\
& \text{Qui bene } \textbf{dansandi} \text{ monstrat habere modum. (Arena [1529] 1758: v. 1–4)} \\
& \text{‘Therefore, he who wants to learn low dances and to quickly become master in the art of dancing, let him make himself learn by heart this book, which shows the method of good dancing.’}
\end{align*}

Because the markup is converted into text formatting, in the above verses anyone familiar with Latin can immediately recognise what is or is not Latin. Such formatting will prove useful further down, as tagging gets increasingly precise and as repeated manual checking of the corpus becomes necessary.

Researchers may wish to distinguish between macaronic words in different languages, as a prospective analysis could hinge on whether the first part of the hybrid is from, for example, Polish or Portuguese. This task can also be automated to a certain degree. Here this is done by relating each poem title to a language\textsuperscript{15} (in Perl, using its numerical index in the series of texts) and then generating the two letters of the lang attribute value, each one representing a language.

\textsuperscript{14} Attribute values for individual languages correspond to the codes given in the ISO-639 language names coding standard.

\textsuperscript{15} The cases of using more than one non-Latin language in a single poem, being extremely rare, are dealt with manually.
Up to this point the process has been almost completely automated. As for the size of the corpus, the only restraining factor is the computer memory. However, much work remains to be done if the objective is tagging various types of words in the macaronic realm. This is the topic of the following subsections.

2.3. Correction of the binary output

The corpus—now tagged in a binary fashion—has to be manually corrected. First of all, some Latin words are not recognised as such by the program; these must be moved to the Latin section manually. Their number and kind depend on the quality of the Latin database at hand and, perhaps contentiously in the present case, on our own decision of what counts as a Latin word. Some common problems in this regard include non-standard spellings (e.g., sydera, lachrymas, michi, terre, foemineus...) and alternate forms (e.g., miraclum, dominabus) which, despite being considered unequivocally Latin by contemporary readers, are often not present in Classical Latin form lists because they only became common in post-classical periods.\(^\text{16}\)

Secondly, the database used by the Morpheus lemmatiser does not include proper names. If one were to generate a form list from an existing word list of Classical Latin that included such names (for example, from a digitisation of Lewis & Short 1891), unless Neo-Latin texts are included as sources for the initial word list—which was not performed in the course of the present analysis, but remains as an option for the future—some non-classical examples which are not recognisable as macaronic, but which are nonetheless familiar to Neo-Latin readers, would not be identified by the program (e.g., Carolus, Franciscus, Avicenna, or Polonia). The same holds for non-classical common nouns that appear so frequently in Neo-Latin that they even find their way into high-style poetry (e.g., taba(c)um ‘tobacco’; paradisus ‘paradise’; Manichaeus ‘Manichean’). In a subsequent manual check, these words were tagged as Latin. Less-common names or ad hoc macaronic creations (e.g., Fransa, Espagnolus, Wrightus), meanwhile, were left among non-Latin words.\(^\text{17}\)

The examples listed above show that Neo-Latin, like every language that has communicated with other linguistic systems, has admitted words from other languages and simply attached Latin endings to them. This process of borrowing is analogous to macaronic vocabulary building, and in fact takes place with any adoption of a foreign word into an

\(^{16}\) I also encountered some unexpected cases such as the absence of forcipe ‘pincers’ (ablative singular), which is certainly a regular Classical Latin word.

\(^{17}\) In deciding the (not unproblematic) status of names in Neo-Latin, I followed handbooks such as Grässe (1972) and Hofmann (1698).
inflectional system of a language; however, in macaronic poetry it is uniquely abrupt and excessive, and readers usually encounter newly-forged words that do not appear outside the macaronic tradition. Such words cannot be considered normal Neo-Latin neologisms, not least because they lack “predictability,” which has been recognised as a reliable criterion for distinguishing code-switching from borrowing.\(^\text{18}\)

There is also some cleanup required on the Latin ‘side’ of the corpus, because the program marks as Latin some macaronic formations that are homonymous with esoteric words from the Classical Latin period. One example is \textit{batone}, a word from a French-Latin macaronic, made up of the French \textit{bâton} ‘a rod’, and an imitation of the Latin ablative ending \textit{-e}, which fits grammatically and semantically into the sentence structure. However, the program recognises this word as coming from the extremely rare Latin word \textit{batus} ‘the blackberry-bush’ or ‘a Hebrew measure for liquids,’ in the ablative case (\textit{bato}), joined with the enclitic \textit{-ne}. It is beyond doubt that the Latin \textit{batus} has never come to macaronic readers’ minds when encountering \textit{batone}. This is why the existence of such words in Classical Latin is discounted form lists.

Given the currently available databases of Latin word forms, both of the above-noted cleanups must be performed manually, i.e., by reading line-by-line and keying in the changes in the attribute values.

2.4. Ternary tagging

In the macaronic tradition, not all non-Latin words are hybrids. Some of them belong completely to the EL. Therefore it is useful to distinguish between hybrid words and pure non-Latin words. Certainly, there is no database of hybrid macaronic words, but for languages that have reliable form lists (which ideally reflect the state of a language at the time a given poem was written), researchers can find pure EL words by way of yet another automated process. On the Internet there are word lists for (the modern variants of) many languages and, in fact, most of the languages appearing in the present corpus are covered.\(^\text{19}\) These lists come in handy at this stage. Here a self-written program is used that first checks for the existence of each word in the applicable word list at appropriate places in the corpus, and then sets the \texttt{type} attribute to the value "p" if the word is found or "h" if it is not. Of course, the fact that the available word lists are not adjusted to the corresponding historical periods and

\(^{19}\) I used lists available at \textit{WinEdt Dictionaries} (2014) .
dialects makes them considerably less appropriate for the present purpose, but the subsequent manual check has shown that they do the job pretty well. This might be partly due to the fact that the pure EL words make up only 6% of the total word count in the corpus.

The program used in this stage is a slightly more complicated (and slower) than the previous one, because it must perform multiple checks across eleven bulky form databases. After another manual cleanup, the corpus reveals an additional set of word groups: pure words in languages other than Latin. Of course, the new situation that this creates can also be visually formatted to represent the three distinct word types.

2.5. Minor word types

After the above described steps the words of the corpus are divided into three groups: Latin, hybrid, and pure EL words, the latter two having been connected to their respective languages. However, the picture is more complicated. All of these groups must be classified into various sub-groups in order for the analysis to be able to provide precise results. This process is very language-specific and, consequently, the most engaging one. Some parts of it can be automated by programming. Others must rely on manual checking because they depend on complex cultural factors that cannot always be precisely predicted or require a rich linguistic documentation—which is still insufficiently available for the historical texts—as to be accomplished computationally. Although the hybrids group contains more entries than any other except for the pure Latin word group, it will be demonstrated below that there are some additional distinctions of word types that can be identified in macaronic language. This is why the type attribute is necessary.

At the beginning of an English macaronic the poet states: “Pandere mens est | Skippantes hinc inde fleas” (‘I have the intention to sing about the fleas that hop to and fro’, Anonymous [1788] 1852: v. 2-3). In this example, the lemmatiser recognised fleas as a Latin word (2nd person singular present subjunctive of fleo ‘cry’), although it is clear from the context that it is either English or hybrid.\(^{20}\) An additional example is patri in a French macaronic, which is not the dative form of the Latin pater ‘father’, but rather an onomatopoeia representing broken glass.

Such words belong to a group that can be termed cryptomacaronic words, which are recognised as Latin by the program, either because of translingual homonymy or due to a glitch in the lemmatiser, but which clearly do not fit that classification when read in context.

\(^{20}\) The structure of the verse suggests that it is hybrid; however, this is not relevant in the present discussion.
These words constitute a special class and must be tagged by hand in the corpus (which is made annotator-friendly through graphical formatting). With a sophisticated and syntactic algorithm for Latin, adjusted to the macaronic situation, this stage may in the future be at least partly automated.

Some cryptomacaronic words are pure EL words (e.g., the English 1st person pronoun *I*, which the program used here recognises as the Latin imperative of *ire* ‘go’); others are hybrid (e.g., the Provençal *balando* ‘dancing’, rather than that word’s meaning in Latin, ‘bleating’). Here these words’ values are set to "pc" (lat. *pura cryptomacaronica*) and "hc" (lat. *hybrida cryptomacaronica*), respectively.

The precise category of cryptomacaronic words is the direct consequence of the corpus processing method in the course of marking it up. If the ELs were subjected to automatic recognition first, lexical overlaps with Latin would be conveniently termed “crypto-Latin”. However, an analysis of a sample has shown that moving the other way round (namely, starting from the ELs), or parallel tagging of Latin and ELs, would require much more manual cleanup. Therefore, starting from Latin, as the most frequent language in the macaronics, is the most economical strategy. In the analyses, cryptomacaronic words will be mostly treated as EL or hybrid words, because they used to be immediately recognised as such by the readers, but a special tag makes the future research of homonymy and its poetic effects possible.

In addition, macaronic authors sometimes played with Latin words without including another language. They mixed stems and endings that do not belong together, but the resulting words should not be considered interlingual (e.g., the future *dicebo* instead of *dicam* ‘I shall say’, or the comparative *vulpior* ‘more fox-like, derived from the noun *vulpes* ‘fox’). Such forms, going unrecognised by the program but still macaronic, are here given the type attribute value "h" while their lang value remains "la".

Hybrid and EL words, detected as such by the computer program or by a human annotator, are certainly distinguishable from basic Latin text, and their salience, producing an effect of “linguistic shock”, is what makes macaronic language special. However, there are word forms that can in a given context belong to more than one of the three types.21 In order

21 The phenomenon of interlingual lexical overlap was used in early modern Spain as an ideological tool; texts composed entirely of words that can be either Latin or Spanish were written. See Woolard & Genovese (2007), who use the term *strategic bivalency*, but also distinguish such practices from the macaronic style (2007: 500), where ambiguity is a byproduct, not a strategy. I learned about their work from a remark by an anonymous reviewer of the present article. Wright (2011: 230) calls such words, as appearing in written texts, *visual diamorphs*. 
to give a more fine-grained distinctions and to enable a research of the role lexical ambiguity plays in the macaronic tradition, it is convenient to include a special set of tags for such words. The fuzziness of interlingual boundaries may have various degrees of representation and importance, depending on the language combined with Latin. For example, Romance languages can be expected to have much more overlap with Latin than Germanic or Slavic languages.

In the present work such words are distributed into the following categories:

- **EL/Hybrid**: words that can be EL or hybrid, but not Latin (e.g., the Spanish *segunda*, ‘second,’ which fits into Latin sentence structure). The type attribute value is "dh". These words can be cryptomacaronic as well, i.e., they can have some other meaning in Latin (e.g., the Italian *testa* ‘head,’ which is also Latin for ‘brick’, ‘shell’ and in later usage ‘skull’). The value for these is "dc".

- **Latin/EL/Hybrid**: words that can fall into any of the three groups (e.g., the Spanish *fortuna* ‘luck’ or ‘destiny’, can have the same meaning in the Latin nominative). The value is "d3".

- **Latin/EL**: words that normally do not receive any ending and consequently cannot be macaronically hybridised (e.g., the French/Latin preposition *de*). These are generally function words that are found in fixed phrases, and very often they are also stop words, which can sometimes be conveniently omitted from searches. The value is "dp".

- **Latin/Hybrid**: this group consists of words like *celebrare* (in Latin, ‘celebrate’, or a hybrid from the Portuguese *celebrar* ‘celebrate’). Although this group theoretically exists, it does not seem plausible to set it apart from ordinary Latin. Such words are not salient enough to break the in-sentence Latin flow, and if they are included as a separate category, every Latin root that has been carried over into another language (Romance or otherwise) would appear in this group. In contrast to the previous two categories, these words are not liable to be perceived as non-Latin. Distinguishing such words from Latin words would only show the number of EL stems derived from Latin, without revealing much about the macaronic nature of the texts.

In the present work, the tagging of cryptomacaronic and ambiguous words was a manual undertaking.
2.6. Word types scheme

As a result of the tasks described above, each word in the corpus is now tagged with one of the 12 different lang attribute values and then again with one of the 8 complex type attribute values (see Table 1 and Table 2).

<table>
<thead>
<tr>
<th>Language</th>
<th>Value</th>
<th>Language</th>
<th>Value</th>
<th>Language</th>
<th>Value</th>
<th>Language</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latin</td>
<td>la</td>
<td>Portuguese</td>
<td>pt</td>
<td>Dutch</td>
<td>nl</td>
<td>Polish</td>
<td>pl</td>
</tr>
<tr>
<td>Italian</td>
<td>it</td>
<td>Occitan</td>
<td>oc</td>
<td>English</td>
<td>en</td>
<td>Czech</td>
<td>cs</td>
</tr>
<tr>
<td>Spanish</td>
<td>es</td>
<td>French</td>
<td>fr</td>
<td>German</td>
<td>ge</td>
<td>Croatian</td>
<td>hr</td>
</tr>
</tbody>
</table>

Table 2. type attribute values.

<table>
<thead>
<tr>
<th>type</th>
<th>value</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>monoglossic (pura)</td>
<td>overt</td>
<td>cryptomacaronic (cryptomacaronica)</td>
<td></td>
</tr>
<tr>
<td>hybrid (hybrida)</td>
<td>h</td>
<td>hc</td>
<td></td>
</tr>
<tr>
<td>EL/hybrid</td>
<td>dh</td>
<td>dh</td>
<td></td>
</tr>
<tr>
<td>ambiguous (dubia)</td>
<td>Latin/EL</td>
<td>dp</td>
<td></td>
</tr>
<tr>
<td>Latin/EL/hybrid</td>
<td>d3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This returns 90 potential attribute combinations (Latin with only "p" and "h", and each of the 11 ELs with each of the 8 type attribute values), together representing a powerful categorising tool for future analyses. Visually formatted after the final annotation stage, the text appears as follows in (3):22

(3) Qui_{dp} volet ergo bassas_{dh} apprendere dansas_{dh}  
Et_{dp} cito dansandi_{h} mestrus_{h} in arte fore,  
Hunc bene de_{dp} testa_{dc} se_{dp} forct_{h} discere librum,

---

22 Latin words are again in the regular font, while non-Latin words are boldfaced. Marks for all word types, except for pure Latin words, are added in subscript. This example contains five word types out of eight.
Qui dp bene dansandiₜ₦ monstrat habere modum. (Arena [1529] 1758: v. 1–4)²³

3. Sample analyses
The following section serves to demonstrate the potential of the corpus, through analysis, to reveal trends in the positioning of various groups of words in macaronic poems. First, the sizes of the clusters of individual word types are analysed. Then, a hypothesis is reviewed regarding whether verse beginnings have a special role in displaying the macaronic nature of the poems.

In the cluster size calculation, the words are divided into three main types: Latin (consisting of [lₐ]p, dp, and d₃ types), hybrid (encompassing h, hc, dh, and dc types), and pure EL words (comprising [non-]lₐp and pc). Minor types have been merged into the dominant types because what is relevant is the reader’s linguistic impression. As Latin is the frame language of macaronic poetry and as everything that is not Latin indicates a peculiar text type to the reader, the objective here is to identify a pattern by which macaronic lexical content interrupts Latin. The most salient distinction exists between the main Latin flow of the text and non-Latin interruptions, which are the points of “linguistic shock” that produce the macaronic effect. Within these interruptions, the readers easily distinguish the hybrids from the pure EL words. Words that are linguistically ambiguous do not contribute to the macaronic effect in the same way as the pure non-Latin lexical units because they are processed as Latin words in the course of cognitive grammatical parsing of the text. Cryptomacaronic words are categorised as the words of the language they really belong to, homonymy not being an obstacle to the understanding of the text. Finally, most of these word types are so infrequent that even their separate inclusion into the analysis (which would make it considerably complicated) would not significantly alter the results.

In the verse beginning analysis, only interruptions of Latin text are what interests us; therefore, the division is binary—Latin words are distinguished from the non-Latin (hybrid and pure EL) lexical material.

3.1. Cluster sizes
Various types of linguistically mixed texts can represent different code-switching patterns. For example, longer stretches of the dominant language discourse can be interrupted by extremely short chunks of heteroglossic elements, or a poem may consist of alternating verses

²³ See the translation in example (2).
in two languages. The annotated corpus of the macaronic texts enables the calculation of the individual same-type cluster sizes and their frequency in each text. In this way, the analyst can obtain an approximate picture of the “granularity” of the mixture, giving information not available by mere word type frequency counts.

Figure 2. Density plots of maximum (A.) and average (B.) cluster sizes in each text, distributed according to the linguistic group of the EL.

Panel A of Figure 2 displays a density plot of the longest clusters in each poem as an indication of the maximum length of a single-language discourse considered acceptable for a macaronic poem. Three curves are generated, representing three linguistic groups of the ELs. The height (i.e., y-axis value) of a particular point of the curve increases as the probability increases that the cluster length at the x-axis under the point is the longest one in a text. Interestingly, in all texts, the longest clusters always consist of Latin words, underscoring the status of Latin as the FL. The figure shows that clusters of over 25 words are very rare, although they appear in all three groups. Apart from that, Germanic macaronic texts seem to be the least prone to larger clusters (dropping to very low densities after the size of 20), while much longer clusters are allowed in the Romance group. For example, at the maximum size of
20 to 25 words, the height of the Romance curve is more than double the height of the Germanic one.

As the longest clusters can be isolated outliers, it is useful a look at the average cluster sizes in individual texts, which can be easily calculated with our corpus. Panel B of Figure 2 shows the density plots for average cluster sizes, again distributed into the three linguistic groups. The picture appears to show some typologically interesting trends. First, it is obvious that the Slavic group has a more dispersed, bimodal distribution. This might be accounted for by the fact that some Slavic authors were educated in Italy, while others remained under the strong influence of German poets. A more detailed analysis could give more precise answers. On the other hand, for the Romance texts, the average lengths are grouped around a little less than three, while for the Germanic ones the central value is lower than two. The considerable difference in the distribution between the three groups suggests that the typological properties of the ELs can be suspected to have influence on the average size of the clusters. An analysis of a larger corpus, with divisions on the level of individual languages, periods, and texts, along with appropriate statistical tests, could reveal precise trends.

---

Figure 3. Frequency of clusters of each size in all texts.

---

24 Isolated words are also included as „clusters“ of length 1.
A considerable discrepancy between the maximal and average cluster lengths suggests that the distribution is not symmetrical. A scatterplot graph representing all cluster lengths in the entire corpus correlated with the frequency of each size (see Figure 3) shows their distribution in the most straightforward fashion. A total of 3,422 clusters is placed on a single graph, in such a way that the darker dots indicate a greater density of values. After relatively high frequencies for one-word “clusters” (i.e., isolated words) the plot exhibits an exponential decay, dropping in number of clusters to fewer than 10 very early on, and then soon down to one or zero. This means that, although Latin has priority in word distribution, the texts tend to be divided into relatively small lexical segments, ensuring frequent switching between Latin and macaronic material.

It would be interesting to compare the macaronic patterns with the rhythm of language switching in various historical bilingual texts, such as those presented in other papers of the present volume. Of course, mere numbers do not suffice. Such comparisons should take into account various levels and types of switching.

3.2. Verse beginnings
Verse is the basic rhythmical unit in poetry. Every verse opening starts a new whole that structures the poem. The beginning of the verse gives flavour to that which follows, and the question arises as to whether macaronic poets sought to “shock” readers with a non-Latin word upfront, or if they began more calmly with a Latin flow to be interrupted later. Therefore, the present analysis compares the percentage of non-Latin words in individual poems with the percentage of verses beginning with a non-Latin word. The corpus markup, as described further above and enriched with XML elements labelling verse, enable the calculation to be performed automatically.
Figure 4. Percentage of non-Latin words at verse beginnings compared to their total percentage in individual works.

The scatterplot graph comparing these two variables (see Figure 4) shows that the percentage of macaronic verse openings is generally positively correlated to the total percentage of macaronic words. The regression line is added to show the direction of the correlation. The correlation coefficient between the two variables ($\rho = 0.62$)\(^{25}\) points to a strong positive correlation, which means that an increase in one variable corresponds to an increase in the other. Moreover, as shown by Figure 5—in which the horizontal width of the shape reflects the frequency of the poems grouping around that particular percentage—non-Latin words are even less frequent at the verse beginning than in the poems taken as a whole. This indicates that not only can no special role be attributed to verse beginnings in achieving the macaronic effect, but also the poets opened their verses more readily in Latin than macaronically.

\(^{25}\) The Spearman correlation coefficient, which is used here for the calculation of the correlation, can take values from -1 to 1; -1 indicates a perfect negative association of ranks (as the variable $x$ grows, the variable $y$ decreases), and 0 means the absence of the correlation, while 1 indicates a perfect positive association of ranks. The strength of the correlation is interpreted in the following way: $0 < \rho < 0.19$: very weak, $0.2 < \rho < 0.39$: weak, $0.40 < \rho < 0.59$: moderate, $0.6 < \rho < 0.79$: strong, and $0.8 < \rho < 1$: very strong.
Figure 5. Beanplot graphs of the distribution of non-Latin words at verse beginnings and in entire poems. The height of the shorter horizontal bars represents the percentages in individual poems.

To make Figure 4 richer and to stimulate thoughts of further investigation, the values are marked with the abbreviated names of the linguistic groups of the ELs (Germanic, Romance, and Slavic) rather than dots. Some overall trends emerge. Germanic texts tend to have a higher percentage of non-Latin words, both at the verse beginnings and in total. Furthermore, most of them are placed marginally with respect to the entire corpus. On the other hand, Romance poems flock relatively close to the regression line and seem to be less dispersed than the other groups.

4. Conclusion
The present paper has had a twofold purpose: to describe the compilation of an annotated digital corpus of macaronic poems, and to demonstrate the usefulness of the corpus in the context of comparative analyses. The increased initial effort and challenging methodological choices required in the first task pay off in the analyses, which are quick and accurate regardless of the corpus size.

The computing tools and procedures described in the present paper can be modified and employed in tagging corpora of other kinds of linguistically mixed texts. Apart from that,
some of the methodological challenges would probably at least partly overlap with those dealt with here. Within the research of macaronics, future development should involve the increase in the size of the corpus as well as the tagging at other levels such as morphologic, syntactic, semantic, and extralinguistic.

The analysis of the cluster sizes has pointed to at least two tendencies. First, genetic linguistic groups of the ELs seem to play a role in the relative frequency of choices between Latin, hybrid, and pure EL lexical material (Figure 2). Second, although longer clusters of monolingual discourse occasionally appear, the increase in the size of clusters corresponds to an exponential decrease in their frequency, which emphasises the high level of granularity of the macaronics in this respect (Figure 3).

The verse beginnings are not more frequently macaronic than would be expected from the overall frequency of macaronic words in individual texts; on the contrary, the two variables are highly correlated (Figure 4). Moreover, the total frequencies of the macaronic words tend to be slightly higher than their frequencies at the verse opening (Figure 5). Apart from that, Figure 4 suggests that at least the texts with Germanic ELs have different distributions than the rest.

One direction of further research would be to compare the frequencies of various word types in different languages, language groups, periods, and works. Precise statistical calculations could reveal whether the differences are related to certain external conditions or whether they may be produced by chance. If significant differences exist, the explanations for them should be researched. The same data can reveal the level of uniformity of the linguistic choices in the entire macaronic tradition. For example, if the frequencies of Latin words in macaronic works are grouped around a central value, this would mean that the authors respected the implicit rule declaring the desirable ratio of Latin versus non-Latin lexical material. A further step would be the comparison of the macaronics with other kinds of linguistically mixed works.

The quantity of multilingual texts from the past that is coming to the attention of the research community will only increase in the future, and the variety of interesting findings will multiply. Such complexity can only be adequately handled by the use of information technology, not only in quantitative research, but also in other types of comparisons. This is the reason it is sound to regard digitisation as a necessary ingredient of every major “next step” in code-switching research.
5. References


Arena, Antonius. [1529] 1758. Introductio ad bassas dansas [Introduction to low dances]. In Ad suos compagnones, qui sunt de persona friantes..., 29–30. London [i.e. Paris]: [s.n.].


Chiesa, Mario. 1972. La tradizione linguistica e letteraria cristiano-medievale nelle “Macaronee”. Giornale Storico della Letteratura Italiana 149. 48–86.


