

Latin Standard Phrases (LSPs) in Blitz Latin

CONTENTS:

1. Why do we need LSPs?	1
2. How are the LSPs found?	4
3. How do we use LSPs?	5
4. Installation of the set of LSPs.	6

1. Why do we need LSPs?

Rule-based and jigsaw translators

There are essentially two ways of creating a modern machine translator. The first is to create a parser that examines each word in a sentence for its grammatical context and assigns a meaning based on that context. This generally provides a 'rule-based' machine translator that can be very accurate grammatically, but provides a stilted, non-flowing translation into English. Ambiguous words with several meanings may be wrongly assigned, a particular problem for a language as ambiguous as Latin. This is the method used by Blitz Latin, where a massive effort has been put into resolving ambiguities.

A major problem for the rule-based translator from Latin into English is that there cannot be a translation which will be both grammatically accurate and of good English readability, since there is no mechanism to effect this. The Latin language deals in concepts (and therefore lacks the definite article 'the' and the indefinite article 'a'), while modern European languages, including English, use precise words. Latin dictionaries can devote columns of space to show how the meaning of a concept such as *lex* ('law') varies according to its context. The professional translations that one sees from Latin to English, made by human experts, have all been fudged in some way in order to render them into proper English. It is for this reason that there are so many expert translations of single Latin texts - no two experts can ever agree on how the text should best be translated!

The second method is that currently used by the creators of machine translators between modern languages. These modern languages are 'rich' languages; that is, they suffer very little from ambiguity and can normally be translated word-for-word from one language to the other. Consequently it is possible to obtain large quantities of texts in one language, which have been translated into English (or other foreign language) by human professional translators. Documents provided for the United Nations and the European Parliament provide fertile sources. These dual language documents can be chopped into small phrases of 1-4 words, with the professional translation of each phrase added as the 'meaning' of the phrase. Longer phrases are preferred, since they provide more context.

Thus the second type of translator, commonly named a 'statistical' translator but much better described as a 'jigsaw' translator, regards every new text as a jigsaw puzzle, to be divided into small phrases, compared with its precomputed jigsaw pieces, and the corresponding professional translations re-inserted. The best way to imagine a jigsaw translator is as a piece of cardboard where the foreign text has been printed on one side, and a professional English translation has been printed on the other side in such a way that the foreign words on one side correspond to English words on the other. If we now chop up the cardboard into lots of small pieces of different shapes and sizes, each with a piece of foreign text on one side and the professional translation on the other, we have the jigsaw pieces that can be used by a translator.

Note that such a jigsaw translator rarely has any concept of vocabulary (its words are character patterns separated by spaces, and replaced by the corresponding character pattern of the professional translation), nor does it have much (if any) concept of grammar. A further difficulty is that different

translators may assign different translations to different jigsaw pieces, while different writers may have used ambiguous phrases to signify different meanings. This is particularly a problem with ambiguous languages such as Latin, but can be partly resolved on a statistical basis: the more occurrences of one translation for a particular jigsaw piece, the more likely it is to be used. Alternative translations can be offered as user-selectable options from a menu.

It is worth emphasising that, unlike the literal translations made by rule-based machine translators, the translations of a jigsaw translator are all rearrangements of existing, beautifully crafted, English prose created by a human professional: wonderfully easy to read, but perhaps not so accurate. Perhaps not accurate at all. Those who cannot read the original language may be impressed by the superb readability of the output, but cannot tell whether what they have read may be a random jumble.

Translations between modern European languages using the jigsaw model are quite surprisingly good, and are championed by such giants of the translation industry as Systran (USA) and SDL (UK). But both now prefer to use a combination of rule-based and jigsaw translations - the jigsaw approach alone has been found to be insufficient for word-for-word translations of even modern languages.

An alternative approach for Latin

Can a jigsaw translator be effective with an ambiguous ancient language like Latin, which has two unusual characteristics (when compared with most modern languages):

1. It is an inflected language, so the ending of each word modifies the meaning and use of the word. For example, 'rex' means 'king', whereas 'regis' means 'of the king', 'regibus' means 'to (or from) the kings' and so on. This requires knowledge of grammar and of context in order to distinguish 'regibus' = 'to the kings' or 'from the kings'.
2. The word order is not the conventional Subject-Verb-Object ('the king loves the queen') of most modern languages, but an order that places the words in order of their intended emphasis - which may vary from one writer to another. That is, the writer could choose to say *rex amat reginam* (literally 'king (subject) loves queen (object)') or use any other ordering. For example, the word order *reginam amat rex* (literally 'queen (object) loves king (subject)') has exactly the same meaning as *rex amat reginam*, but with emphasis that it is the queen who is loved, and the reader is presumed to understand the inverted word order by reading the inflections of each word.

One can be fairly sure that a jigsaw translator created, for example, by chopping up professional translations of the Vulgate Bible and the collected works of Cicero is likely to be very effective at translating the Vulgate Bible and the collected works of Cicero. It is difficult to imagine that it will fare so well with writers with other styles, or with medieval texts, or with modern Vatican texts. Even if the writers intend the same phrases as those found in Cicero, they may vary the word order of the Latin in order to change the emphasis. In short, once a jigsaw translator is taken out of its comfort zone, from which its jigsaw 'pieces' were created, it becomes rapidly less reliable. Moreover, the Latin language is particularly susceptible to the construction of compound words, where a new word, that will be easily recognisable to the experienced reader, is created by adding an appropriate prefix or suffix to an existing Latin word. These compound words will not be recognised by a jigsaw translator. By contrast, a rule-based translator can be made to recognise compound words, and has no 'comfort zone', since its function is independent of any other Latin text. However, its translation is much less readable than that of a jigsaw translator.

Is there a better route for translation from Latin into English, that combines the accuracy of a rule-based translator with the improved English wording of a jigsaw translator? We believe that there is, and over six months of programmers' time has been put into implementing the alternative with Blitz Latin (from version 2.0).

The central idea is to create the finest rule-based Latin translator possible, but to overlay its output with precompiled 'Latin Standard Phrases', which replace the worst examples of poor wording (or mis-translation) from the original English output. The application of the Latin Standard Phrases depends very much on the Latin words that have been finally agreed from ambiguous text by the underlying rule-based translator - unlike the case with a jigsaw translator, the replacement English phrase will only be used if it makes sense according to the rules.

Latin Standard Phrases (LSPs)

Standard phrases are not well adapted for use with an inflected language such as Latin. For example, the single verb *amare* (to love) has about 180 inflected variants, incorporating such forms as *amo* (I love), *amabas* (you [singular] were loving), *amabit* (he will love), *amavistis* (you [plural] have loved), *amarantur* (they might have been loved), to cite only a small selection. If we wish to incorporate a LSP such as *amare reginam* (to love the queen), then the phrase will have to be added 180 times if it is to catch every possible alternative. It is true that a statistical search of real Latin texts will determine that far fewer than 180 variations of this LSP actually exist, but who is to say whether other forms might not exist in a wider set of Latin files?

A particular complication for Latin is that the word order of a Latin sentence does not follow the conventional Subject-Verb-Object ordering that is so familiar to users of modern European languages. In the case of the simple example *amare reginam* cited above, it is in fact more likely that the phrase will be encountered in the form *reginam amare*, so that it will be necessary to add another 180 variations of this LSP in order to cover all possibilities. In particular, the emphasis of words that is required by one writer of Latin may be inverted by another writer of the same phrase, according to the latter's idea of which words need to be emphasised.

Conjugation of verbs

Careful coding enables the verb to be picked out in a short Latin clause such as *amabas reginam*, even when it is not the first verb in the clause. Then, knowing the grammatical structure of the verb from the original Latin text, it is possible to determine that the infinitive form is *amare*, and hence the LSP *amare reginam* should be used. Thus Blitz Latin now contains the standard phrase 'scire fatalia/to know something of fate', which embraces options as diverse as *scisse fatalia* ('to have known something of fate') and *scio fatalia* ('I know something of fate.')

Similarly, nouns and adjectives also have inflections that require separation, typically twelve for a noun and 36 for an adjective. To add to the confusion, many of these inflections are re-used, further adding to the ambiguity of translating Latin text. It is much harder to connect these nouns and adjectives to their use in a LSP, and at present Blitz Latin does not attempt this. Thus, Blitz Latin can translate the various inflections of the verb in the LSP *amare reginam*, where the accusative form of the noun *regina* is used, but it cannot assign correctly a similar LSP *amare regina* if encountered as *amare reginam*.

Misspellings in Latin

It is customary for modern machine translators to search simply in order to match jigsaw pieces to the text to be translated. However, Latin is an ancient language, written originally in the days before dictionaries were widespread and hence liable to misspellings. During medieval times, many words became spelled phonetically, rather than according to classical example. For example, the Latin word *gratia* (grace) might be spelled as *gracia*. During the same period, the letters 'j' and 'v', unknown to classical Latin except on inscriptions (when 'v' is easier to carve than 'u'), became widespread. Moreover, the speakers of Latin were as lazy in their speech as any modern teenager, and contracted words, or the inflections on words, to shorter forms for their own convenience. A simple example is the Latin word *ille* ('that one', 'that thing', the precursor of the definite articles *il/la* in modern Italian), which is shortened repeatedly in early Roman plays to *ill'*.

It is evident that any attempt to apply LSPs to Latin text with these imperfections is likely to give unreliable results. Blitz Latin, however, has long enjoyed the ability to correct all these errors and to store the corrected word within its data structure. Thus, for Blitz Latin, LSPs are compared with the corrected forms of the Latin, and not with the original Latin text.

2. How are the LSPs found?

Blitz Latin now incorporates more than 3,100 *published* LSPs. The majority have been taken from Carl Meissner's personal list, published in English in 1894 from the 6th edition of the original German text. Meissner's words are heavily biased towards the use of phrases with infinitives (eg *fortuna secundi uti* - 'to be fortunate') which are particularly useful with Blitz Latin's coding for infinitive verbs.

Meissner's phrases possess the defect that they are taken exclusively from classical Latin. In order to fill the gaps for medieval and neo-Latin, the dictionaries of Lewis and Short ("A Latin Dictionary", OUP, 2002 reprint), Collins ("Latin Dictionary plus Grammar", HarperCollins Publishers, 1997), Stelten ("Dictionary of Ecclesiastical Latin", Hendrickson Publishers, 1995) and Latham ("Revised Medieval Word List", British Academy, 1965) have all been consulted and suitable short phrases adapted for use with Blitz Latin, while a large number of user phrases from the author's personal collection (acquired while puzzling over some of the less comprehensible translations from Blitz Latin) and from other minor sources have also been included.

An alternative method for adding Latin Standard Phrases to Blitz Latin is to make a fast computer search of an existing set of Latin files. The author's test collection of Latin files includes the Packard Humanities Institute selection, used with permission, and containing all known Latin texts up to about 200 AD, together with a few later files. Other sources have provided almost every other Latin text known to the author from later Latin (up to about 500 AD), so that classical and late Latin are covered very well. The 1,500 years of Latin writing that have elapsed since then are covered in a haphazard way that reflects broadly the medieval and neo-Latin files that others have found sufficiently interesting to type up and post on the Internet. A few have been sent in by users of Blitz Latin, a confidential procedure that ensures that Latin words hitherto unknown to its electronic dictionary are likely to be included in future.

At the time of writing, the total number of LSPs in Blitz Latin exceeds 15,800. The obvious criticism that might be made of the method of collecting LSPs from a real collection of Latin files is that the phrases will be bounded by the Latin found in the test set. Other users might have different Latin collections. Users of Blitz Latin can also add their own LSPs to the supplied long list. These additions are subject to certain rules, see section (3).

Another difficulty may be mentioned in connection with the statistical process for collecting LSPs. Latin authors in ancient and (especially) in medieval times showed a marked tendency to copy verbatim chunks out of each others' texts. For example, difficult Latin phrases from the poet Vergil were discussed by later writers, asking what the poet actually meant. Vergil's poetry was constrained by the need to fit words to his favourite meters, and therefore later writers quoted several of his phrases verbatim in order to discuss them. Thus it happens that what was originally just one idiosyncratic phrase becomes magnified into a popular 'standard phrase'. A large number of medieval writers quoted Boemus, who wrote a long treatise on the subject of music theory. Again, large chunks of Boemus' work are quoted verbatim. And again, Christian writers, especially St Augustine, repeatedly quote large chunks out of the Latin Bible. It is difficult to know how best to handle these potential LSPs, since on the one hand a single rare phrase has been copied into prominence. On the other hand, undoubtedly the phrase now occurs frequently, spread across several later writers. The current policy is to include these much-quoted phrases as LSPs, with the important exception of phrases taken originally from the poets where the phrases are artificial constructs anyway.

The problem of clashing LSPs, where for example the first and second words form one standard phrase, and the second and third Latin words could form another LSP, is resolved on the basis of first-come-first-served, and the internal counter to test Latin words is advanced to the end of the first LSP. To supply a contrived example, the Latin clause *ita ut quid...* contains the LSPs *ita ut* ('just as') and *ut quid* ('to what purpose'). Blitz Latin's translation is therefore 'just as what' and not 'just as to what purpose'.

3. How do we use LSPs?

Effect of LSPs on Translation Speed

A characteristic of Blitz Latin is that it is intended to execute very quickly, as the name suggests. Therefore a key issue with the coding for Blitz Latin of the LSPs remains a preoccupation with the effect on the run-time speed for translation of real Latin texts.

Application to the LSP code of the same AI principles used to search Latin stems resulted in the time overhead being reduced to just 0.6% on the standard set of PHI files of classical Latin. A considerable expansion of the LSPs supplied with Blitz Latin should have little further effect on program translation speed.

Instructions for Users

Menu options Edit/User Phrases/Allowed and Edit/User Phrases/Show permit the user respectively to switch any use of the LSPs ON or OFF, or whether or not to show the encapsulating characters <...> (see below).

The LSPs of Blitz Latin are contained in a single text file, entitled 'UserPhrases.txt'. All users of Blitz Latin will receive a very basic set of files in the provided original file 'Userphrases.txt'. However, registered users will receive a file that is (currently) around 300 times larger. The entries are pre-sorted into alphabetical order for the convenience of users, but the latter can add their own LSPs freely to the bottom of the text file. [NOTE: In order to protect intellectual property, most registered users will receive encrypted files of Latin Standard Phrases. However, such users can still add their own further LSPs in unencrypted form to the bottom of the file, providing the rules stated below are still met. Encrypted LSPs are distinguished by a leading asterisk '*'.] The only limit on the number of LSPs added to Blitz Latin is the free computer memory (RAM) available when the program attempts to load the LSPs - the quantity of such free RAM depends heavily on whether memory-intensive computer programs are simultaneously in use, and a warning will be given if the list of LSPs has been truncated owing to insufficient RAM.

The structure of the LSPs in the file is exemplified thus:

a retro/behind\in arrears

The forward slash ("/") is essential, and separates the Latin phrase being sought from the English translation. The back-slash ("\") is entirely optional and serves only to separate alternative translations. Very few of the LSPs used in Blitz Latin require such alternatives.

The only other requirements are these:

Maximum five Latin words in the LSP, each of a maximum of 20 characters.

Maximum 78 characters in the English meaning.

Any use of a Latin verb infinitive requires identification of the infinitive (INF) in the Latin phrase and of the word 'to' (TO) of the verb in the English meaning. The infinitives are identified by their position in the phrase, counting the first word as '1'. For example:

a re publica recedere/INF4TO1to retire from public life

where recedere is the 4th word of the Latin phrase, and 'to' (of 'to retire', the verb) is the 1st word of the English meaning. This usage distinguishes 'to' pertaining to the verb from 'to' as in 'to the dog'.

Other examples:

ad te scribere/INF3TO1to write to you

nisi opus esse/INF3TO3unless there to be need

An extension of these LSPs (from version 2.07) permits the use of phrases comprising two word stems that match one another for case, gender and single/plural. One of these two words must be a noun, the other an adjective or participle. For example:

reg. amat./the loved king

The word *reg.* denotes the stem for *rex, regis*, king; the word *amat.* denotes the stem for the participle *amatus/a/um*. The periods (full-stops) are essential.

Introduction of the two-word phrases permits a wide range of matched noun-adjective combinations to be given meanings. For example, if the text file *UserPhrases.txt* contains the example given above, any occurrence in the Latin text of *regem amatam* will be translated as '<the loved king>', while any occurrence of *regis amati* will be translated as '<of the loved king>'. Note that *rex amatus* will not be converted into a LSP, since it lacks one of the required two stems.

A noun-adjective combination of this sort cannot be combined with a verb in the same Latin Standard Phrase. There must be a presumption that the verb will control any noun present; for example, a transitive verb will require an object noun to be present.

A further issue is that of Latin word order. The LSPs *mens. ligne.* and *reg. amant.* will assign their standard phrases to *mensae ligneae* and to *regem amantem*, but not to *lignae mensae* and *amantem regem*.

It is particularly important to recognise that Blitz Latin applies its User Phrases to Latin texts AFTER the original Latin words have been parsed. The words within Blitz Latin may sometimes be different internally from that which you thought you had typed.

Two examples from user complaints:

1. '*gens meroingorum*'. *meroingorum* is not recognised as a Latin word by Blitz Latin – it is not in the electronic dictionary. LSPs cannot be expected from words that are missing from the electronic dictionary. However, if we use the word '*Meroingorum*' (first letter of the word is now upper-case) the word is recognised by Blitz Latin as a declineable proper name by default. Therefore, if the original text had been written as '*gens Meroingorum*' the LSP would have functioned correctly. LSPs are case-indifferent, but they cannot be applied to words where the translator has failed to recognise one of the words.
2. '*appropinquasset*.' This is a subtle problem. '*appropinquasset*' is a contracted form (known as a syncope) of the real spelling '*appropinquavisset*'. Blitz Latin recognises this, and converts the contracted form into the full form internally, and invisible to the user. Thus, what is required is a LSP for '*appropinquavisset*', and not for '*appropinquasset*'.

The use of a LSP in Blitz Latin is marked with a preceding "<" and a trailing ">", for example:
a re publica recessi. <I have retired from public life>.

These enclosing characters can be switched on and off at will by the user.

All the Latin words must be provided in lower-case characters only. The translation program sorts out whether or not words with upper-case characters in the original text can be matched to the LSP.

Likewise, Blitz Latin is indifferent to whether the user has provided Latin phrases with the letters 'u' or 'v', or with 'i' or 'j'. These can be jumbled up indiscriminately, and the program will convert all into 'u' and 'i', as used internally.

A more awkward problem concerns the use of enclitic words (*-que, -ve, -ne*), exemplified by the short Latin phrase *rex reginaque amant suum canem* ('the king and the queen love their dog'). To the machine translator, there is no difference between this phrase and the similar *rex et regina amant suum canem*. However, the standard phrase for *rex reginaque* must be entered as *rex que regina*, because that is how Blitz Latin handles the text internally.

Unwanted single LSPs can be switched off by the simple expedient of placing a double hyphen "--" in front of the offending line in the text file of LSPs.

4. Installation of the set of LSPs.

The LSPs will be supplied together with the supplementary medieval dictionary, in a package for registered (paying) users only.

1. The package will be supplied in a self-installing, compressed file, meddict.exe. Place this file in a convenient folder on your C: drive (folder C:\temp will do very nicely). Only the C: drive will serve for this purpose; however your folder for Blitz Latin does not itself have to be in the C: drive.
2. Double-click on the file meddict.exe to 'unzip' (expand) the two files contained into the same folder. In the example, they will be unzipped to give files C:\temp\meddict.dat and C:\temp\userphrases.txt.
3. Now run your copy of Blitz Latin (version 2.0 or above). It will search for the two files and copy them across into the correct folder automatically. Or, of course, you may choose to move the files yourself to the correct folder yourself. Locate your file 'blitz2.exe' and use Windows Explorer to drag across the files meddict.dat and userphrases.txt into the same folder where Blitz2.exe resides. [Note to Apple users: Blitz Latin was not designed for use with Apple computers. However, several users have informed us that Blitz Latin functions correctly with Apple computers fitted with a Windows emulator and Microsoft Windows XP (the 'Wine' emulator seems to give the best results.) In your case, it is essential to drag the two files with the mouse into the correct folder, since the Apple computer lacks a 'C:' drive.]
4. Once the two files have been located and copied by Blitz Latin (or installed manually), then Blitz Latin will continue to operate normally, and you will see these two messages in the right-hand window of the screen:
Loading medieval dictionary...Done.
...
Loading Latin Standard Phrases...Done.
5. And that finishes the installation task! Happy translating!