A LETTERS JOURNAL EXPLORING THE FRONTIERS OF PHYSICS

OFFPRINT

# Horizontal transfers are a primary aspect of languages evolution

MICHELE PASQUINI and MAURIZIO SERVA

EPL, **125** (2019) 38002

epl

A Letters Journal Exploring the Frontiers of Physics

# AN INVITATION TO SUBMIT YOUR WORK

epljournal.org

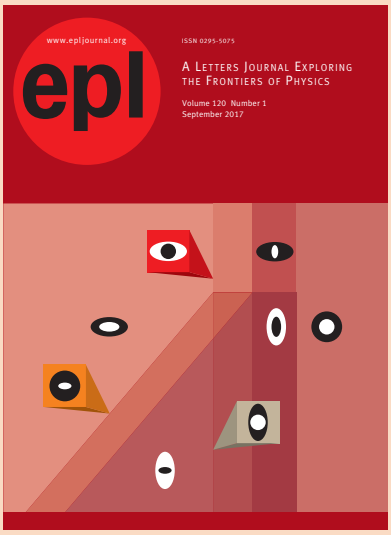
## The Editorial Board invites you to submit your Letters to EPL

Choose EPL, and you'll be published alongside original, innovative Letters in all areas of physics. The broad scope of the journal means your work will be read by researchers in a variety of fields; from condensed matter, to statistical physics, plasma and fusion sciences, astrophysics, and more.

Not only that, but your work will be accessible immediately in over 3,300 institutions worldwide. And thanks to EPL's green open access policy you can make it available to everyone on your institutional repository after just 12 months.

## Run by active scientists, for scientists

Your work will be read by a member of our active and international Editorial Board, led by Bart Van Tiggelen. Plus, any profits made by EPL go back into the societies that own it, meaning your work will support outreach, education, and innovation in physics worldwide.


**OVER**

**638,000**

**full-text downloads in 2017**

**Average submission to online publication**

**100 DAYS**

**21,500**

**citations in 2016**

*We greatly appreciate the efficient, professional and rapid processing of our paper by your team.*

**Cong Lin**
Shanghai University

## Four good reasons to publish with EPL

1. **International reach** – more than 3,300 institutions have access to EPL globally, enabling your work to be read by your peers in more than 90 countries.

2. **Exceptional peer review** – your paper will be handled by one of the 60+ co-editors, who are experts in their fields. They oversee the entire peer-review process, from selection of the referees to making all final acceptance decisions.

3. **Fast publication** – you will receive a quick and efficient service; the median time from submission to acceptance is 75 days, with an additional 20 days from acceptance to online publication.

4. **Green and gold open access** – your Letter in EPL will be published on a green open access basis. If you are required to publish using gold open access, we also offer this service for a one-off author payment. The Article Processing Charge (APC) is currently €1,400.

Details on preparing, submitting and tracking the progress of your manuscript from submission to acceptance are available on the EPL submission website, **epletters.net**.

If you would like further information about our author service or EPL in general, please visit **epljournal.org** or e-mail us at **info@epljournal.org**.

**EPL is published in partnership with:**

European Physical Society

Società Italiana di Fisica

EDP Sciences

IOP Publishing

epl A Letters Journal Exploring the Frontiers of Physics

February 2019

EPL, **125** (2019) 38002 www.epljournal.org
doi: 10.1209/0295-5075/125/38002

# Horizontal transfers are a primary aspect of languages evolution

Michele Pasquini and Maurizio Serva

*Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica, Università dell'Aquila - L'Aquila, Italy*

received 8 November 2018; accepted in final form 28 January 2019
published online 5 March 2019

PACS 89.75.Fb – Structures and organization in complex systems
PACS 89.75.Hc – Networks and genealogical trees

**Abstract** – More than sixty years ago Morris Swadesh proposed to measure the similarity between two languages by considering their overlap, *i.e.*, the percentage of cognates. The overlap was estimated by comparing two *ad hoc* lists with words corresponding to same meanings for the two languages. The notable assumption of Swadesh was that replacements in a vocabulary occur at a universal constant rate so that the time distance from the eventual common ancestor can be determined. In his view only replacements are relevant while horizontal transfers (borrowings from other languages) are less important and their effect can be eventually taken into account assuming that the divergence is diminished by contact. Later, the mainstream of glottochronology adopted the point of view that the effect of loanwords could be eliminated by careful work devoted to their identification so that they would not affect any measure of distance between languages. The aim of this paper is to show by experimental evidence that horizontal transfers, on the contrary, are a primary aspect of languages evolution since their effect on the vocabulary is at least as important as that of spontaneous replacements. We finally show that this phenomenon severely and unavoidably limits the possibility to fully reconstruct a proto-language. This limitation is fundamental, *i.e.*, it gives a bound which cannot be infringed, independently of the method used for the reconstruction.



**Introduction and material.** – According to Swadesh [1–3] each word has a constant probability rate $\alpha$ to be replaced so that the percentage of words of a given language which remain unchanged in a time $t$ is $\exp(-\alpha t)$. If two languages have a common ancestor $T$ years in the past, this assumption allows to determine $T$ from the overlap $C$ which is the percentage of cognates (words with the same etymological origin). In fact, given that the words in common are those which remained unchanged in both languages, it is easy to derive $C = \exp(-2\alpha T)$ which implies $T = -\frac{1}{2\alpha}\ln C$.

The dynamics induced by replacements is the only phenomenon which is considered relevant while borrowings from other languages are considered somehow unimportant. Indeed, Swadesh himself [2,4] tried to take into account the effect of horizontal transfer by a phenomenological assumption that contact diminishes the separation between two languages. His proposal (rewritten in our notation) was that the formula which gives the time from the common ancestor should be modified as $T = -\frac{1}{2\alpha s}\ln C$, where $s \leq 1$ is a parameter which takes into account the effect of diminished divergence due to contact. Nevertheless, $s$ is not the output of a process which incorporates borrowings but it simply accounts for the fact that horizontal transfers make languages divergence slower. Moreover, by principle, the parameter $s$ in not measurable from data, only the product $s\alpha$ can be computed if $C$ and $T$ are known.

This fact was clear to Swadesh, who also states his fundamental formula as a bound on $T$, *i.e.*, $T \geq -\frac{1}{2\alpha}\ln C$, where the equality holds only in the absence of horizontal transfers. Finally, we mention that Swadesh himself did not attempt to use this formula for anything. In a much more recent paper [5], another parameter was used to quantify the effect of borrowings on the divergence between languages, a parameter which has the advantage that in specific circumstances (historically attested events) it can be eventually measured but which is still phenomenological.

Later work of glottochronology abandoned Swadesh's attempt and tried to treat horizontal transfers as some kind of *noise* that could be cleansed by a careful work devoted to the identification of loanwords [6–9]. A subsequent deletion of their contribution from all metrics should totally neutralize their effect on languages comparisons. Roughly speaking, the standard glottochronological

approach reduces the lexical evolution to a strictly vertical process.

Our view is definitely different since we think that horizontal transfers are a primary aspect of languages evolution. In order to support our point, we will try to provide experimental evidence that the effect of borrowings on the vocabulary is comparable to that of spontaneous replacements. With this respect, our work can be viewed as a realization of Swadesh's proposals for incorporating the effect of horizontal transfers.

In order to argue for our thesis, we consider the case of Latin and contemporary Romance languages. The reason for our choice is twofold: the ancestor language (Latin) is perfectly known and the daughter languages (Romance languages) are numerous and precisely recorded.

Our main source for Romance and Latin languages was "The Global Lexicostatistical Database, Indo-European family: Romance group" (September 2016 version), which contains annotated Swadesh lists of 110 meanings compiled by Mikhail Saenko. This database is part of the "The Global Lexicostatistical Database", which can be consulted at `http://starling.rinet.ru/new100/main.htm`.

In this database two versions of Latin are considered: Archaic Classical Latin (ACL in the following) and Late Classical Latin (200–300 CE). Since it is commonly accepted that Late Classical Latin is not at the root of Romance languages, but rather it is a side branch of the spoken Latin, we consider ACL. ACL is based on the plays of Titus Maccius Plautus (254–184 BCE, born in Sarsina). In difficult cases the vocabulary is implemented by data from "*De agri cultura*" by Marcus Porcius Cato (234–149 BCE, born in Tusculum). Thus, ACL can be situated approximately 2200 years before present.

Probably, the last common ancestor of all contemporary languages was the Latin commonly spoken in the first centuries of the current era (1500–2000 years before present), a more recent date with respect to the 2200 years of ACL. Nevertheless, ACL is for sure a common ancestor of all of them even if not the last common ancestor.

The dataset we use also contains 55 contemporary Romance languages, the complete list can be found at the bottom of fig. 1. An exhaustive description of all our dataset, which also contains a description of our additions to the "Global Lexicostatistical Database" (many words in various languages and two entirely new languages), can be found in the appendix of [10].

Let us also mention that in this paper we use an automatic approach to compute the overlaps [5,11], this is without consequences since everything can be re-obtained by the traditional method of Swadesh of subjective cognate identification. The way we compute the overlap $C$ is explained in detail in the first section of a recent paper [10] where it is also compared with the standard cognate procedure.

**Framework and results.** – According to the glottochronology assumption that replacements occur at a universal constant rate $\alpha$, the overlap between a contemporary language and its ancestor language spoken $T_L$ years before is $C_L = \exp(-\alpha T_L)$. In fact, $C_L$ is simply the percentage of words that did not change in the daughter language during its evolution from the ancestor language.

The overlap between two contemporary daughter languages which started to differentiate $T$ years before present from a common ancestor is instead $C = \exp(-2\alpha T)$. The factor 2 at the exponent in this second case is explained by the fact that both daughter languages undergo vocabulary replacements at a rate $\alpha$ and $C$ is the percentage of words that did not change in both languages.

In our case, the parameter $T_L$ is the known time distance between ACL and each of the contemporary Romance languages, therefore, $T_L = 2200$ years. Moreover, a reasonable assumption is that regional varieties of Latin began to differentiate as soon as Latin started to be spoken outside Rome (more 2150 years before present) and that differentiation became a ubiquitous phenomenon at the end of the western empire (1550 years before present). Therefore, we should have for $T$, depending on the pair of languages, something between 1550 and 2150 years.

In order to find out the inconsistency of the main stream view it is useful to recast the problem in a differential form. According to the universal replacement rate hypothesis, the overlap $C_L$ can be derived by the simple differential equation below on the left:

$$\dot{C}_L = -\alpha C_L \to C_L = \exp(-\alpha T_L) \to \alpha = -\frac{1}{T_L}\ln C_L, \tag{1}$$

where the initial condition $C_L = 1$ leads to the solution of the differential equation (center) which, in turn, leads to the equality on the right.

The value of $C_L(i)$ can be computed from lexical data for each Romance language $i$ as well the corresponding $\alpha(i)$, given that $T_L = 2200$. An average of $\alpha(i)$ can be done in order to obtain a more reliable value of $\alpha$. This average can be done over all the $M = 55$ Romance languages (see the bottom of fig. 1) or a smaller number. Besides $M = 55$, we also considered a set of $M = 15$ languages scattered uniformly over all the Romance linguistic area (the first 15 languages at the bottom of fig. 1), a set which only contains $M = 4$ languages (Standard Portuguese, Standard French, Standard Italian and Romanian) at the four sides of the linguistic area and, finally, a set composed only by Standard Portuguese and Romanian ($M = 2$) which are the two languages with a greater geographical separation.

The result can be found in table 1. It can be easily seen that the value of $\alpha$ is largely independent of the number $M$ implying that all languages evolved from ACL with a similar rate of replacement whose value is about $\alpha = 4.4 \times 10^{-4}$ ($\alpha$ is roughly the probability that a given world changes in one year). It should be noticed that borrowings play no role here and the result is by principle the

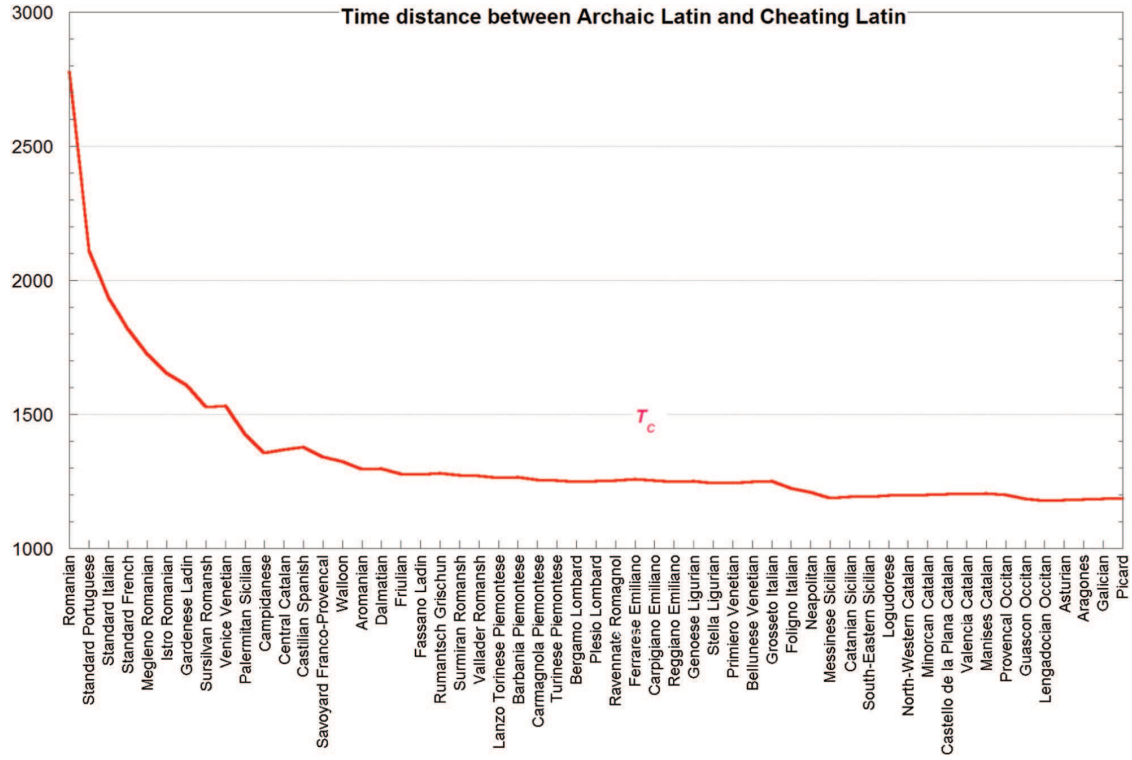



Fig. 1: Time distance $T_C$ between ACL and CL computed by using an increasing number $1 \leq M \leq 55$ of Romance languages (inserted following the order at the bottom of the figure). The thermalization occurs at a value of $T_C$ close to 1200, the same we obtained for the time distance between the virtual ancestor language and ACL.

Table 1: Parameters $\alpha$, $T_W$ and $\beta$ computed for different subsets ($M = 2, 4, 15, 55$) of Romance languages.

| | $M = 2$ | $M = 4$ | $M = 15$ | $M = 55$ |
|---|---|---|---|---|
| $\alpha$ | $(4.3 \pm 0.5) \times 10^{-4}$ | $(4.2 \pm 0.6) \times 10^{-4}$ | $(4.5 \pm 0.6) \times 10^{-4}$ | $(4.4 \pm 0.5) \times 10^{-4}$ |
| $T_W$ | $(1.12 \pm 0.13) \times 10^{3}$ | $(1.02 \pm 0.03) \times 10^{3}$ | $(1.03 \pm 0.08) \times 10^{3}$ | $(0.89 \pm 0.11) \times 10^{3}$ |
| $\beta$ | $(2.3 \pm 0.4) \times 10^{-4}$ | $(2.8 \pm 0.1) \times 10^{-4}$ | $(2.6 \pm 0.4) \times 10^{-4}$ | $(3.5 \pm 0.7) \times 10^{-4}$ |

same, independently of the presence or absence of horizontal transfers. This can be simply understood because a borrowing replaces a word with another one which has the same probability of coming directly from ACL.

Again, according to the standard glottochronological approach, the evolution of the overlap $C$ between two Romance languages is governed by the differential equation (left)

$$\dot{C} = -2\alpha C \rightarrow C = \exp(-2\alpha T_W) \rightarrow T_W = -\frac{1}{2\alpha} \ln C, \tag{2}$$

where the initial condition $C = 1$ leads to the solution in the center.

Given that all values $\alpha(i)$ have been already computed by eq. (1) and having defined $\ln C(i)$ as the average of $\ln C(i,j)$ over all other languages $j \neq i$, we have $T_W(i) = -\frac{1}{2\alpha(i)} \ln C(i)$ which can be computed for each language $i$. The average of $T_W(i)$ over $M$ languages $i$ can be also found in table 1.

In all cases we have that $T_W \simeq 1000$ years which is incompatible with $T$ whose expected value ranges between 1550 and 2150. Notice that this value of $T_W$ also comes out when one only considers Portuguese and Romanian ($M = 2$) which, as a matter of fact, started to differentiate more than 1000 years ago.

In conclusion, we have to admit that something is wrong in the standard glottochronological description.

Our suggestion is that Romance languages shape a fully connected network in which horizontal transfers cannot be considered as accidental but they are a constitutive aspect of the phenomenology. Regional variants started to differentiate from the beginning but they remained closer than expected just because of horizontal transfers.

According to our view, the equation on the left of (2) must be replaced by the equation on the left below:

$$\dot{C} = -2\alpha C + 2\beta(1 - C) \rightarrow$$
$$C = \frac{\beta}{\alpha + \beta} + \frac{\alpha}{\alpha + \beta} \exp(-2(\alpha + \beta)T), \tag{3}$$

where the parameter $\beta$ accounts for the possibility that two different words (with same meaning) in two different languages become equal as the effect of a borrowing. The time from the common ancestor is chosen for all pairs to be equal to $T$ (1550–2150 years before present, let us say for simplicity 1850).

With the initial condition $C = 1$ the above equation can be solved (right). Moreover, given that all values $\alpha(i)$ have been already computed by eq. (1) and having defined $\ln C(i)$ as before, these two values can be inserted in (3) and we can get $\beta(i)$ by solving numerically for any $i$. Finally, averaging over the $M$ languages we get for $\beta$ the results shown in table 1.

Formula (3) and the formula of Swadesh $C = \exp(-2\alpha s T)$ are very different and the latter simply states that time separation is larger of what expected in the absence of borrowings but is useless in practice. In fact, while $\beta$ and $\alpha$ can be both computed from data, the Swadesh formula only allows the computation of the product $\alpha s$.

The important output is that the value of $\beta$ is comparable with the value of $\alpha$. This is significant especially when we consider $M = 2$ and $M = 4$, in this case, in fact, the choice $T \geq 1850$ is reasonable since the four languages started to differentiate at least two-thousand years before present.

Since the effect of horizontal transfers is comparable with the effect of spontaneous replacements, the phenomenon cannot be considered accidental or negligible, but rather it constitutes a primary motor of language evolution.

Moreover, the differences in the value of $\beta$ that we see for the different values of $M$ have a simple explanation: for larger $M$ two compared languages are on average geographically closer. Since closer languages more easily exchange words, we can safely assume that the effect of borrowings becomes more and more visible when $M$ increases.

Therefore, the weakness of the model eq. (3) is that it is a mean field, *i.e.*, the value of $\beta$ is independent of the geographical distance of two languages. A more realistic model should include geography, so that this spurious $\beta$ dependence on $M$ should disappear. Nevertheless, this difficult task is well beyond our goal which is simply to show the relevance of horizontal transfers.

In conclusion, we think that we have provided enough evidence for our thesis, in fact, even in the case $M = 2$ the value of $\beta$ is more than one half of the typical value of $\alpha$, which means that even Romanian and Portuguese cannot be considered languages with a separate evolution.

**Reconstruction of a proto-language.** – We would now to discuss the relevance of our result for what concerns the reconstruction of a proto-language.

According to our proposal, $T_W = -\frac{1}{2\alpha}\ln C \simeq 1000$ years measures the time distance of modern Romance languages from their *virtual* ancestor language. It should be clear that $T_W$ is not the time when they started to differentiate, but rather a time which measures their mutual difference *as if* they evolved separately.

In other words, assume you know $\alpha$ and you do not know that horizontal transfers operate, you would conclude that the common ancestor of modern Romance languages was situated about one-thousand years in the past. We know, however, that it was situated at a time $T$ which is about twice $T_W$.

The time $T_W$ is given by an average over all meanings since the time depths of each word can be different. For example, the Italian word *ciao* which is spread (with different spellings) in all Romance area (indeed worldwide), originated from the Venetian *s'ciavo* (slave) about 200 years ago, while other Italian words as *animale* or *mano* can be linked directly to ACL (2200 years from present).

Since the virtual ancestor language is situated $T_W = 1000$ years before present and not $T \simeq 1850$ years, we are prompted to conclude that some of the proto-language words (which existed $T$ years ago) disappeared from all Romance area and that they were replaced by more recent common ancestors words. It is very unlikely that these disappearances of ancestral words were caused by independent replacement in all the languages (almost impossible if the languages are about 50 and if they do not communicate). But it is reasonable to assume that there were horizontal transfers causing a single word, which originated in a single language, to replace everywhere the corresponding Latin (or Vulgar) word (as it was for *ciao*). In this case the proto-language word is lost forever and it cannot be reconstructed by looking at the present languages.

We can easily conclude that the best we can do while attempting to reconstruct a proto-language from its daughter languages is to obtain the virtual ancestor language (if reconstruction is optimal).

Reconstructing the proto-language, on the contrary, is impossible since many words are lost forever because of borrowings.

The time distance between present Romance languages and virtual ancestor is $T_W$, therefore, the time distance between the virtual ancestor language and the true proto-language is $T - T_W$ while the time distance between the virtual ancestor language and ACL is $T_L - T_W$.

If we do the best possible (optimal reconstruction) we obtain a reconstructed language which coincides with the virtual one and, therefore, it is situated $T_L - T_W$ years from ACL. If our reconstruction is sub-optimal we obtain a reconstructed language that is even more different from ACL so that its time distance from it is larger than $T_L - T_W$. Therefore, our result concerning the distance between the reconstructed language and ACL should be $T_L - T_W \simeq 1200$ (for an optimal reconstruction) or larger (for a sub-optimal reconstruction).

In order to have an optimal reconstruction we cheat, in fact, we use our knowledge of ACL, while in the relevant cases a common ancestor language is unknown (for example for Indo-European languages). We will describe the cheating procedure in a while (let us call this reproduced

language "Cheating Latin") but we anticipate that the procedure is conceived in a way that it is impossible to have better imitations of ACL using Romance languages. Let us also anticipate that our results confirm all previous results since they provide a time distance between Cheating Latin (CL in the following) and ACL of about 1200 years which equals $T_L - T_W$.

In this way we have a twofold result:

- our previous estimate $T_W \simeq 1000$ years is confirmed by an independent procedure;
- even if we cheat we obtain a language (CL) which is separated from ACL by $T_L - T_W \simeq 1200$ years and, therefore, cannot be the proto-language which is separated from ACL by $T_L - T \simeq 350$ years.

In other words, at our best we obtain the virtual ancestor which is macroscopically separated from the proto-language.

As expected, when we tried to reconstruct the proto-language without cheating (by various procedures which only involve contemporary Romance variants) we obtained results which are even worse, *i.e.*, they are at a larger time distance from ACL (about 1500 years).

Let us describe our cheating procedure. CL is a language constructed taking, among all the words with same meaning in a given set of $M$ Romance languages, that word which is most similar (in the NLD metric described in [5]) to the corresponding ACL word. Therefore, by definition, it is the best imitation of ACL which can be created only using modern Romance languages.

In order to measure the overlap between CL and ACL we have to take into account the bias which is generated by casual matching between words. In order to explain this point consider the case in which we try to reconstruct ACL following the same procedure of CL but using Romance words whose meaning does not correspond to the meaning of the ACL word. For example, to reconstruct *manus* (*hand*) we may consider the Romance words corresponding to *big* and we find *mannu* which, in Logodurese (Sardinia) means *big*. Not so bad, *manus* and *mannu* are quite similar, but this similarity is accidental. Let us call Wrong Latin the language constructed with this procedure.

To avoid this bias we first construct CL and we measure its overlap $\bar{C}_C$ with ACL, then we compute the overlap between Wrong Latin and ACL. Indeed, there are several different ways for matching two different meanings and our result is obtained by an average over these possibilities. In this way we obtain $C_A$ that we may call accidental overlap (the overlap between Wrong Latin and ACL). Finally, the overlap between CL and ACL, after elimination of bias, is $C_C = (\bar{C}_C - C_A)/(1 - C_A)$.

The unbiased overlap $C_C$ can be computed using any number of Romance languages. Forcefully, the more languages, the better the reproduction. In fig. 1 we computed the time distance $T_C = -\frac{1}{\alpha} \ln C_C$ using an increasing number of Romance languages. It can be seen that there is an asymptotic-like behavior corresponding to 1200 years that is exactly the same we obtained for $T_L - T_W$. This means that the best we can do for Latin is to find out a reconstructed language which can be temporally situated around the year 1000 CE.

As we already wrote above, we also constructed several proto-languages by using only contemporary Romance languages (completely ignoring ACL). In this case the reproduction is forcefully worse and the time distance from ACL is larger (we obtained about 1500 years for the best cases instead of 1200 years). In conclusion, because of borrowings the proto-language cannot be reasonably fully reconstructed. This result does not exclude the possibility of a reconstruction in those cases in which all or some of the daughter languages remained separated.

**Discussion.** – We hope that we have provided enough qualitative evidence that horizontal transfers have a primary role in languages evolution. It is quite strange that linguists have not taken up Swadesh's suggestion from 1955 and have ignored hidden borrowings, concentrating only on those borrowings whose phonological shapes have allowed for their identification. In our opinion, at least in the case of fully connected networks such as the Romance geographical area, the relevance of the horizontal transfer phenomenon is almost self-evident.

In our formulation we give a measure of $\beta$ which, roughly speaking, is the probability that a borrowing of a word corresponding to a given meaning occurs in a year. As already mentioned this quantity is affected by geography and a more realistic evaluation of $\beta$ should emerge from a more detailed model which includes spatial distances and physical barriers (seas, mountains, ...) as fundamental ingredients. Moreover, in order to use $\beta$ for quantifying the number of borrowings in a given time some hypotheses should be done on the underlying dynamics (borrowings may be from one language to another or two languages may borrow the same word from a third one, ...). Both the inclusion of geography and the detailed description of the nature of the borrowings are well beyond the scope of this paper, but if one could succeed in this research, eventually, the amount of borrowings estimated from our model could be compared with previous estimates. For example, in [11] some new estimates are provided based on the results of the Loanword Typology project and also some previous estimates are also reported.

We expect to find that our estimated rate of loanwords is larger than previous estimates based on direct observation. This difference would be mostly due to loanwords among closely related language varieties that cannot be detected because the inherited words have not been differentiated enough. To give an example: English has two words with the same ultimate origin: *shirt* and *skirt*. The latter was borrowed from Norse, and this can be detected because *skirt* was borrowed after *sk* became *sh* and this

sound change ceased to operate. But if the borrowing had happened when *sk* was still in operation the borrowing would have been too subtle to detect. The original meaning of the inherited Germanic word was something like a short garment. That developed into the meaning *shirt* in English and *skirt* in Old Norse. If Old English had borrowed the word before the $sk \to sh$ change it would inherit, the fact that English would have changed the meaning of the word from a garment worn on the upper part of the body to one worn on the lower part would be too subtle to enable the linguist to identify this as a borrowing.

Moreover, it may happen that same new word appeared almost at the same time in two geographically close languages. In most cases it is almost impossible to decide in which language the new word appeared as a replacement and in which language it was imported. Probably, we should also accept that this question is almost irrelevant.

In conclusion, there are many words that linguists cannot identify as having been borrowed, the amount of this hidden borrowing could emerge from a comparison of the result of this paper with previous estimates, but, for the reasons we exposed previously, we are unable to accomplish this task here and we simply make the proposal to investigate this topic in a situation where everything can be quantified easily and exactly.

Let us also mention that the presence of horizontal transfers, which more likely occur between geographically closer languages, has an interesting application. It is, in fact, possible to reconstruct geography only by the set of overlaps between pairs of languages [12]. It is as if one could draw the chart of Spain only speaking about football with people from different places of that country sitting at the same table.

A side effect of our findings is that proto-languages cannot be really reconstructed. Even if a word is shared by most of the languages of a contemporary family, there is a non-vanishing probability that it originated somewhere after the proto-language disappeared and that it later diffused everywhere by borrowings (by contacts, like a disease). This phenomenon severely and unavoidably limits the possibility to fully reconstruct a proto-language. This limitation is fundamental, *i.e.*, there is a bound which cannot be infringed, independently of the method used for the reconstruction.

In contrast, historical linguists will argue that one can find the proto-language of a family of related contemporary languages. The case of Latin and its offspring is really useful for testing this kind of statement since not only the modern Romance languages are well recorded but also a language (Latin) closely related to the proto-Romance is known. In this case one should be able to find something which was spoken in the first five centuries of the current era, which means a distance of 200–700 years from ACL. On the contrary, our cheating procedure (the best possible) leads to a reconstructed language (CL) which has a distance of 1200 years from ACL. This result is consistent with our model which, taking into account horizontal transfers (3), provides the same distance.

We think that the case of Romance languages and Latin (or Vulgar) could provide a good opportunity to test the reliability of specific recipes for proto-language reconstruction. This would help to evaluate the degree of precision of the recipe even in those cases where a direct test is impossible (as, for example, proto-Indo-European reconstruction).

In the future, we would like to apply our ideas to Malagasy dialects, for which we have a large dataset [13] that we collected a few years ago and which we are updating and extending. In this case, the goal will be to obtain a reconstructed language situated in the past with respect to contemporary dialects, to be more fruitfully compared with Indonesian languages.

REFERENCES


[1] Swadesh M., *Int. J. Am. Linguist.*, **16** (1950) 157.
[2] Swadesh M., *Southwest. J. Anthropol.*, **7** (1951) 1.
[3] Swadesh M., *Proc. Am. Philos. Soc.*, **96** (1952) 452.
[4] Swadesh M., *Int. J. Am. Linguist.*, **21** (1955) 121.
[5] Serva M. and Petroni F., *EPL*, **81** (2008) 68005.
[6] Lees R. B., *Language*, **29** (1953) 113.
[7] Gudschinsky S. C., *Word*, **12** (1956) 175.
[8] Starostin S., *Comparative-historical linguistics and lexicostatistics*, in *Historical Linguistics & Lexicostatistics* (Association for the History of Language, Melbourne) 1999, pp. 3–50
[9] Starostin G., *J. Lang. Relationship*, **3** (2010) 79116.
[10] Pasquini M. and Serva M., *Stability of meanings versus rate of replacement of words: an experimental test*, arXiv:1802.06764v2.
[11] Holman E. W., Wichmann S., Brown C. H., Velupillai V., Müller A. and Bakker D., *Folia Linguistica*, **42** (2008) 331.
[12] Serva M., Vergni D., Volchenkov D. and Vulpiani A., *EPL*, **118** (2017) 48003.
[13] Serva M., Petroni F., Volchenkov D. and Wichmann S., *J. R. Soc. Interface*, **9** (2012) 54.