

Chapter 2: Armenian Spell Checker

The fool rolled a stone into a pit; a hundred wise men could not draw it out.
While the prudent man considers; the fool is across the river and away.
Replies of the fool are the proverbs of the people.

HySpell, The Remedy

In 1989, my life experienced a turning point, while I completed my Ph.D. in Pure Mathematics at UCLA, the professorship careers were at their lowest in the history of United States. My mother, after suffering for twenty years from chronic heart condition, had just passed away two years ago, and the memory of her in my heart was still as fresh as days away. The challenge of not being able to get a teaching or research position had already inflicted a misery upon my personality. As if all these were not enough, a flu epidemic spread in the town, and somehow, I was exposed to it. The flu infected my left inner ear, and while I was ill in bed with a timpani beating inside my left ear, it occurred to me that what I was seeking for so many years in revealing the secrets of the Universe are very much dependent on the primitive concepts of one's mother tongue. So I decided to relearn my entire 25 years of mathematical study and research but this time write my studies entirely in Armenian. This grand program revived my life and strengthened the purpose of my existence, and before soon I was on my Mac II Classic box to start typing the first few pages of my studies using the Armenian fonts that I myself had designed for the T_EX (the famous technical typesetting application of Donald Knuth).

This romantic love was not left undisturbed, back then in Armenia the political and economic turmoil on one hand and the devastating earthquake on the other, had taken their toll on my beloved Armenia, and my heart could not rest thousands of miles away while the mountains of my ancestors seem to be calling my name in unison.

To make the long story short, I got a job as software engineer with the purpose to get back to my grand programme as soon as the call for my duty is accomplished and my financial conditions improve. However, I did not know back then that this decision would put 18 years of my life in limbo. So, here we are, just like my forefathers Khorenatsi and Shiragetsi, after 20 years of my undesired career as software architect (understand it as "in exile"), I am back to the starting point of my mathematics/physics grand programme, and to my surprise, people have not yet developed a practical Armenian spell checker tool for Microsoft Word. We have an abundance of Armenian fonts; "Yes, thank you Lord...", but where is the beef? Shame on Armenian software developers of the world, and yes, shame on me too.

I guess, twenty years of unwanted professional experience in software development was worth living. It finally paid its dues, and soon, it was time for me to develop HySpell, the first practical Armenian spell checker for Mesrobian orthography inside Microsoft Office Word 2007. I came to learn however, even with my linguistic background, developing HySpell was not a piece of cake. It ate almost two years of my life, and finally on November 2009, I began to see the light at the end of the long tunnel.

Mastering a language

What exactly do we mean by “mastering” a language? To answer this question, let us consider the linguistic skills of a toddler, eight year old, teenager, high-school graduate, college graduate, amateur writer, and finally a professional writer. These linguistic skills are characterized by several factors of which the vocabulary-base is one of the most fundamental factors. The larger your vocabulary base, the more colorful and creative your writing skills become. On the other hand, the larger the vocabulary base the more difficult to remember them all. Therefore, there is a limit to how large you can attain and maintain your vocabulary base, be this in speech or writing. In most cases, people who don’t write at all have much less vocabulary base than people who often write. Some people have a habit of singing songs, which does improve their vocabulary base along with their speech skills. Incidentally, the habit of singing songs is the most influential method of learning or in some cases assimilation. So, if you have this habit and you have reason to believe that you are being assimilated, then stop singing like a broken record, or even better, sing Armenian songs instead, so that you can at least appreciate this assimilation.

The more phonetic a written language is, the easier the spelling becomes. However, the written forms of most languages do not correspond in a one to one phonetic fashion with the spoken forms. Classical Armenian was perhaps one of the most phonetic languages of the families of ancient languages (if it was not the only one). Nevertheless, because of numerous dialects that developed over long periods of isolation from each other, the near perfect phonetic qualities have been lost in some of these dialects. This is especially the case in the Western Armenian dialect. Therefore, the Armenian language spell checker especially for a Western Armenian speaker certainly is an indispensable tool.

A spell checker, besides pointing and identifying misspelled words in the text, also, has a list of suggested words for possible replacement or correction. This function in most circumstances improves the ability of a writer by simplifying some of the otherwise tedious proofing tasks.

Most proofing tools also possess a thesaurus or electronic dictionary, with which the task of a writer becomes considerably manageable and practical. Although in the current introductory version of HySpell, we have not included such a dictionary, this functionality will certainly be added in the subsequent versions. The thesaurus will include functionality of finding Armenian synonyms, antonyms and homophones to a given Armenian word, as well as the descriptive definition of the word in Armenian and English (i.e. Armenian-English dictionary). Note that by homophone we imply the same pronunciation but different spelling and different meaning, for example, յարկ (floor) and հարկ (taxation). Recall that in regards to spelling, the homophones are a more serious problem for a Western Armenian dialect speaker, especially the generations that emerged from the ashes of the 1915 Genocide, and who settled in Middle East, France and USA. Most of the Western

Armenian sub-dialects prior to this period had somehow preserved the proper phonetics of the Classical Armenian, as it was spoken during the Cilician Armenian Kingdom. Evidence of such may be found in the Mousa-Ler, Moush, Tigranakert, Van and other Armenian sub-dialects. This pronunciation variation of the Armenian alphabets originated most likely from the Constantinople Armenian sub-dialect, which was heavily influenced by the Byzantine and later by the Ottoman-Turkish languages. And since around 1880's this sub-dialect, after a gradual purification process, became the official language that was taught in the region, it eventually dominated over the other spoken Western Armenian sub-dialects.

Mesrobian Orthography

Why HySpell implements only Mesrobian orthography? My answer to this is perhaps already implied by the chronicles of the fool (i.e. read the notes right below this chapter's heading). Perhaps I should elaborate on this. Well, we all know that if I did not do so, there will soon emerge a huge scandal on my head. Nevertheless, this is not my reason. My honest reason is: because, I am an advocate of the Classical Armenian language, and a true follower of what Khachatour Abovian said in his famous novel "Վէրք Հայաստանի" (Wounds of Armenia). I believe that not only Diaspora is very much capable to renaissance back to the spoken and written Armenian language, but it can go even further and learn Classical Armenian. And taking the initiative to learn the old way of writing will shorten this path and thus, achieve my main goal.

The advocates of what is know as 1920's Soviet reform of the Armenian orthography, will try to paint a chaotic picture that, around 1920's Soviet Armenia had to take a strict measure to reform the Armenian language because, historically this process was under continuous struggle ever since 12 century and therefore, being in such an unstable condition, this reform was extremely critical and necessary. This is not quite true, because at the time there was already a well formed literature and both the Eastern and Western Armenian intellectuals used a single and well-formed orthography, namely the classical orthography. Evidence of this may be found in the original publications of all writers in the era, as well as, from the works of linguist H. Acharian, S. Malkhasiantz and others. In fact, it took another 20 years for the official acceptance of a modified form of the reform orthography, and it was quite painful transition period wasting a lot of writers' efforts (e.g. read the complains of Acharian, Malkhasiantz, Toumanian and others to get a true historical picture). In contrast, my "act of the fool" will not be so painful, because automaton will do the laborious work, while the master of automaton enjoys the fruits of her/his imagination.

On the other hand, the advocates of the classical Mesrobian orthography are so antagonistic towards the 1922 reform, that 80 years of excellent literature is discriminated solely on the basis that they are printed in the 1922-reform orthography. In some cases, this discrimination is so much polarized that poetry of Toumanian and Shiraz are irresponsibly altered into Western Armenian during the conversion of the orthography. I consider such an act a barbarous act. It makes me so sad and angry to read Toumanian and Shiraz in such altered form, in the same way, I would become sad and angry when I read Tekeyan's poetry in Eastern Armenian.

Here is my approach to the problem:

1. First of all, Western Armenian and Eastern Armenian are not different languages. They are only subsets of the same language, namely the Armenian language.
2. Secondly, there are many more sub-dialects that are currently in use inside Armenia proper, and therefore, the particular choice of the dialect is a choice for the writer (e.g. Toumanian's work is a good example of such success).
3. A gradual return of orthography to the classical orthography is an ideal strategic step. The main reason for this is uniformization of the written form of the language without lose of linguistic evolutionary root formation. This later aspect is in fact much more important than the phonetic correspondence with the alphabet. Finally, the process of returning to classical orthography is no longer a difficult or tedious process in this age of computers, and HySpell spell checker can be an indispensable tool for this purpose.
4. The phonetic pronunciation of the Western Armenian must gradually be corrected back to the classical Armenian phonetics similar to that of the Eastern Armenian. This is perhaps a more difficult task, but it is not impossible if mnemonic alphabet songs are developed and properly taught to children at an early age. Although in this case, a child may still habitually use the wrong pronunciation during conversation, but at least knowing how to pronounce the alphabet correctly is half way to solving the problem. A lot of western Armenians can easily speak eastern Armenian, and in most cases with the correct pronunciation. So, this problem will in most probability be solved by itself through time.
5. Elementary schools in Diaspora and Armenian must encourage the use of a finalized version of classical orthography presenting literature repertoire that includes works of all Armenian writers without tampering their dialectic structure (i.e. eastern, western, as well as other dialects). There is nothing difficult or confusing in this requirement. I, for one, am a product of such a system, and I grew to love Armenian literature. I can speak and write in several dialects including eastern, western, classic, Lori, Moush, Kanaker and others.
6. By finalized version of classical orthography, I mean the classical orthography that preserves all the root word orthography based on Malkhasiantz or the 19th century orthography prior to 1922 reforms, along with some minor consistencies (which are mostly related to hyphenation of certain patterns, and will be decided in subsequent releases of HySpell).
7. Along with this orthography issues some classic or dialectical phonetic techniques must be introduced. In other words, this century has an advance technology compared to previous centuries (i.e. the times have changed), and therefore this technology must be utilized to arrest the speech deviation evolutionary process of the language, just like the written form of the language arrests the dialectic diversity evolution of the language. This is not an act against the natural evolution of the Armenian language. On the contrary, the technology is carelessly being incorporate to abolish the boundaries of this natural process, and the only way to fight against it is to use the same weapon against this assimilation.

In conclusion, I do not see any problem here for actual resolution of these issues. All that is needed is the correct and well thought foundation and framework, under which all Armenian dialects by their own free usage may evolve into a uniform and purified standard.

The phonetic factor of written language indeed simplifies greatly its use, but there are limitations to this factor, and that eventually symbolism dominates the more conceptual aspect of the language especially during its real-time communication. This is why there were many conceptual symbols still used in manuscripts written after the 5th century (some of them borrowed from the pagan era). A

more up to date example could be the several hundred symbols and notations used in today's mathematics and physics. In other words, when it comes to complex concepts and their real-time communications, symbolism is much more powerful than linear phonetic-alphabetic text, assuming of course that the actual conceptual meaning of the symbolism is already known. This later fact is the main reason why so many language re-phonetization reforms were doomed to failure. To convince the reader about this, try reading the following *Bobdot*-phonetic English text:

Hé red uböot the síklòn which had awlmòst hit the sivik sentur. The skí had lûkt tpretning. Hé gáv a sí. Ther wûd hav bin the devul tú pá if the storm had bin wurs. The nolij mád him wins. Hé t̄pawt, "Wé ar só helples in the fás uv bad wethur." Hé gáv sum t̄pawt tú sólor powur, but klung tú hiz yúzhúul rigur, which ment discüsing this wit̄p hiz wíf bèföör máking a dèsižhun. Shé ekspèkted this cunsiduráshun. Hé had bin t̄piñking uböot èmürjensè powur and had an ìdéuh. Sum mezhur uv akshun mít bé prúduunt.

Or try reading the following English text in another phonetic method:

If th feurst pair ov vouulz in eny seeqns iz wun ov th 13 pairz ov vouulz abuv, then it iz oenly spoekn az abuv, and if foloed bie 1 aur 2 maur vouulz, th feurst pair iz spoekn feurst. Faur egzampl [poeit] (poet), [dueet] (duet). bt if th feurst 2 vouulz aar not wun ov theez 13 diegrafs, then th feurst vouul iz spoekn sepreutly. Heer aar sum komn multipl vouulz.

Incidentally, the above phonetic English methods are probably the easiest of the several such English orthography reforms. The reader may also attempt to transliterate the above text (or any English text) using the Armenian alphabets instead of Latin to get an idea.

Taking this modern era English phonetic reform sample texts in mind, and after a considerable amount of time thinking about both classical and modern orthography, it is not difficult to realize that their difference is so negligible that it was not worth for 1922 reform to cause such frustration that hampered the unity of the Armenian nation. And what is more ironic is that this unnecessary reform was later blamed on Manuk Abeghian, an intellect that is probably one of the greatest masters of Classical Armenian and Armeneology. My true belief is that if Manuk Abeghian was asked about such a reform under scientific free consciousness (and not the Soviet order), he would have probably laughed at your face, with the following response: "Do we want to introduce phonetic reform on the most phonetic language on Earth? This is unheard of..."

In this regards, therefore, unlike some Diaspora intellectuals, I do not want to alienate Manuk Abeghian as the author for the Stalinist plot against the Armenian language and nation. After all, Manuk Abeghian was a great Armenian scholar, and indeed at the end, the result of the language reform was cleverly minimized to the most negligible amount as I pointed out in the previous paragraph. I believe that Abeghian was trying to defend and reverse the already controversial and erroneous developments in orthography that were initiated by Ghazaros Aghayan and exploited by others. And therefore, this "negligible amount" was deliberately devised by Abeghian to minimize the damage caused and to avoid further conflict with Soviet decret, in this way, once for all ending the reform issue.

Here is what H. Acharian has written about Ghazaros Aghayan in regards to V-sound issue –

“Այսպէս, ուրեմն, Ղ. Աղայեանի փաստերը ամբողջապէս սխալ են եւ ամենեւին ո՛չ մի բանի զօրութիւն չունին: Հեղինակը մի յետին միտք ունէր. այն է՝ իբր թէ բարեփոխել հայերէնի ուղղագրութիւնը եւ նրա միջից երեք տեսակ V գրութեան ձեւերից (ւ, ու, վ) երկուսը ջնջելով՝ պահել մէկ հնչիւնին մէկ գիր: Նպատակը բարի էր, բայց միջոցը սխալ իսկ արդիւնքը բաբելոնական խառնակութիւն, որի մէջ ընկաւ այն ժամանակ արեւելեան հայերէնի ուղղագրութիւնը Աղայեանի պատճառով:” [p. 481, Հայոց Գրերը, Հ. Աճառեան 1984].

Finally, advocates that perceive that a language is nothing but a tool for communication are deeply mistaken. For if this was the case, then why do we have so many languages on the planet instead of a globalized single universal language. The answer to this later question is the underlying strategic essence, which St. Mesrob Mashtots, many centuries ago implied when he invented separate alphabets for the Georgian and Aghvank languages instead of incorporating the already existing Armenian alphabet. This essence is indeed the only sword that can be used against the assimilation of the new twenty-first century’s juggernaut called “Globalization”.

Given all these aspects about orthography, they do not constitute the least problem of the practical usage of the Armenian language in Diaspora, and lately, even inside Armenia proper. In United States and France, the latest generations of Armenians have already lost the usage of the Armenian language (even at speech level), while on Internet the Latinization of slang Armenian is creeping in every news commentary and blogs. The scientific communities in Armenia are more and more publishing their research papers and articles in English and Russian, without their corresponding Armenian translations. With this rate, it is a matter of only a few decades, during which, the Armenian language will become a dead language practiced only by archeologists and historians.

It is therefore this extremely critical matter that drove me to the creation of HySpell site.

Using HySpell

The HySpell Armenian spell checker for Microsoft Office 2007 Word is integrated side-by-side with the native spelling and grammar command toolbar in the Word application, and is accessible via the **Armenian Spelling** ribbon-button in the HySpell ribbon-region (see Figure 1 for more details).

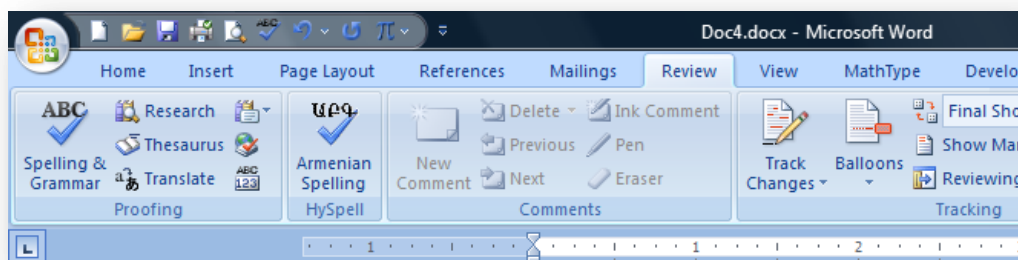


Figure 1. HySpell’s Armenian Spelling ribbon-button

To spell-check Armenian text in your document, simply open your Word document and click the **ԱԲԳ Armenian Spelling** ribbon-button to run the Armenian spell checker program. This will initiate

the spell checking process and will proof-read the Armenian text paragraph by paragraph. If it meets any misspelled word, the spell-checker's dialog will be displayed informing the end-user the found mistake and its position in the document. Figure 2 below shows this dialog in action along with all its visual details.

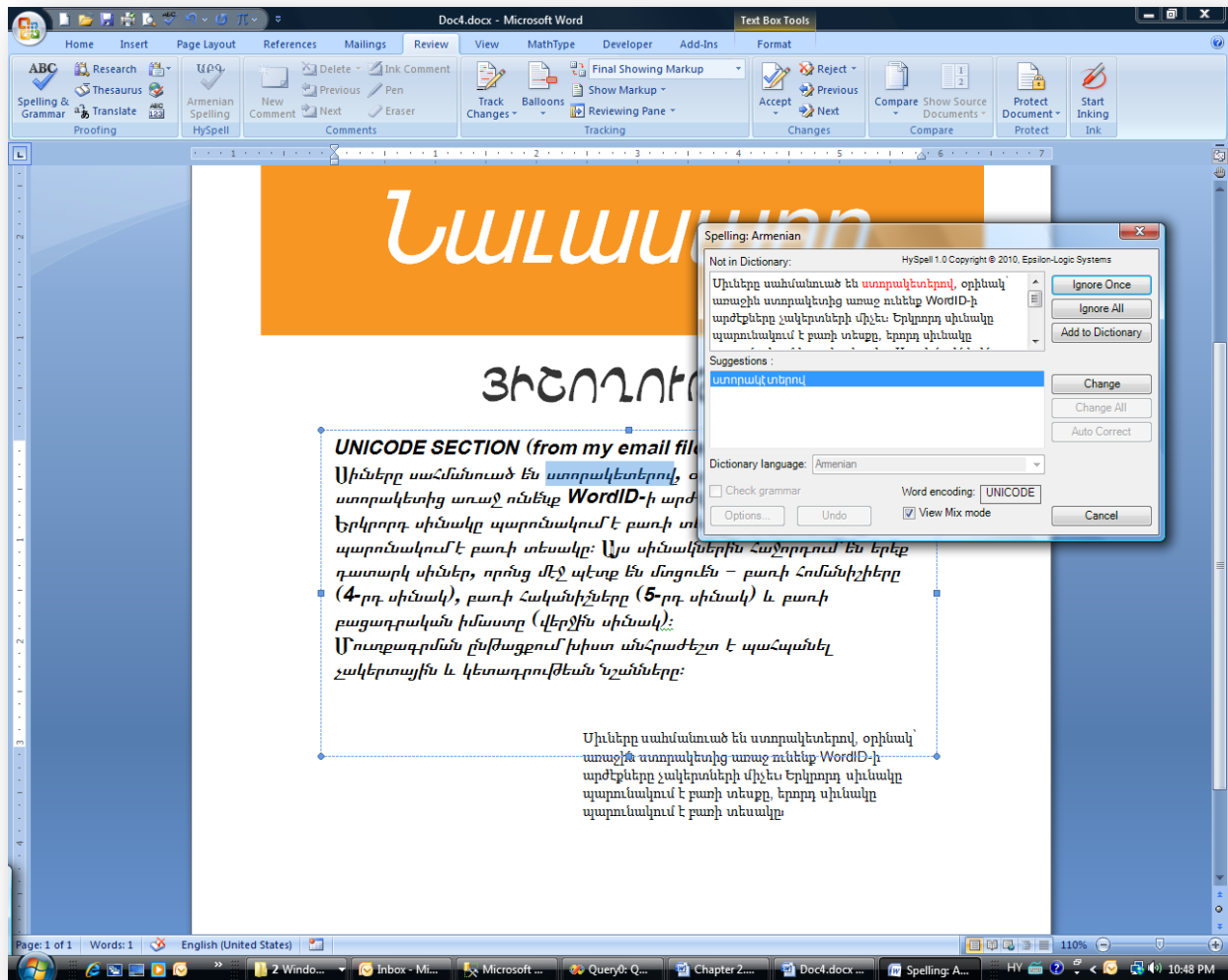


Figure 2. Showing the HySpell Armenian spell-checker's dialog in action

In particular, the dialog consists of the following fields and command buttons:

Not in Dictionary	text-box	<i>in which, the misspelled word is displayed in red inside the text of the processing paragraph.</i>
Suggestions	list-box	<i>in which, a list of legal words are suggested for the end-user to select and correct the misspelled word.</i>
Dictionary language	combo-box	(disabled in version 1.0)
Check grammar	check-box	(disabled in version 1.0)
Word encoding	display-field	<i>this field displays the encoding of the processing word (possible values are UNICODE or ARMSCII-8).</i>
View Mix mode	check-box	<i>this check-box will alternate the encoding mode of</i>

		<i>the Not in Dictionary text-box, and by default it is set to mix-mode (i.e. displaying both UNICODE and ARMSCII-8 mixed encoded text).</i>
Options	button	(disabled in version 1.0)
Undo	button	(disabled in version 1.0)
Ignore Once	button	<i>clicking this button will ignore the currently selected misspelled word and proceed to the next possible misspelling.</i>
Ignore All	button	<i>clicking this button will ignore all occurrences of the currently selected misspelled word in the rest of the text of the document. It will then proceed to the next possible misspelling.</i>
Add to Dictionary	button	<i>clicking this button will add the currently selected red colored word of the Not in Dictionary text-box into the custom dictionary. This feature can be used to add new words, such as proper names, abbreviations and other tokens to the spell checkers dictionary.</i>
Change	button	<i>clicking this button will correct the misspelled word in the document by the selected suggestion item. Note that double-clicking the selected suggested word item in the Suggestions list will essentially perform this same action.</i>
Change All	button	(disabled in version 1.0)
Auto Correct	button	(disabled in version 1.0)
Cancel	button	<i>clicking this button will close the spell checkers dialog window and terminate the spell checking process.</i>

There are a few limitations that the user must keep in mind when using this preliminary version 1.0 of HySpell Armenian spell checker product. They are as follows:

- (1) In this 1.0 version, the Armenian spell checker process will only proof read the Armenian text in the document. So, if you have a multilingual document, you will also need to spell check your document for the other non-Armenian text using the Microsoft Word's **ABC Spelling & Grammar** ribbon-button. Subsequent versions of HySpell will include the proof reading of the non-Armenian text along with the Armenian, so that the user will need to use only one spell check command to proof read the entire document.
- (2) HySpell supports only the major encoding for the Armenian text. These are the UNICODE and the old ARMSCII-8 encoding. If you have an old document that does not use these standard Armenian encoding, you will need to convert your text into one of these supported encoding prior to proof reading your document via HySpell spell checker. For example, some schools have an encoding of the Arasan Armenian font type in circulation that uses the old 7-bit ASCII encoding space of the English language. The font designer of Arasan, most probably had the keyboard nuisance in mind when assigning this erroneous encoding, so that, the end user may use a single keyboard driver for both English and Armenian. This erroneous encoding has already backfired on the community of Armenian

users, and HySpell not supporting it is yet another proof that playing with standards is not a good apprenticeship. It is my vehement advice to all Armenian software developers to strictly adhere to the Armenian Standards that are derived from the international standards organizations, such as UNICODE. Any deviation from such standards will in fact backfire on us in the future (if it has not yet done so).

- (3) In version 1.0, we still have the red misspell marking (i.e. underlined in red wiggled line) of all Armenian text in the document displayed. In this case, Microsoft Word is assuming that the Armenian text in the document is in English (or in the OS local) language and therefore indicating that it is misspelled. Our advice regarding this anomaly is: to not depend on this misspell marking feature of Microsoft Word, and instead, to always use the **ABC Spelling & Grammar** or **ԱԲԳ Armenian Spelling** ribbon-buttons to proof read your document. In any case, this Microsoft Word feature is helpful only when your document contains limited amount of misspelled words or other tokens that do not exist in the Microsoft Office's dictionaries, and whenever a document exceeds this limit, the misspell red marking feature will automatically be disabled. In most cases, this does happen when your document passes 30 or 50 pages (even when it is entirely in English).
- (4) The **Add to Dictionary** feature of HySpell spell checker, will add the selected missing word in the HySpell's custom dictionary. This is a good feature for adding new tokens, such as proper names, abbreviations and other word segments, but it does not add any inflected form of the given word. Armenian language being a very flexible inflectional language (unlike English) certainly needs an inflectional add-word feature. This feature will be added in the subsequent versions of HySpell software. As for now you simply may need to add more tokens.
- (5) Regarding inflectional add-word feature, in the current version, there is a way to add a token in inflected form, but this method is manual and requires a little more knowledge on the end-user's side. If the need for such an add-word feature becomes very critical for you, you may contact support@hyspell.com for further information or instructions.

Finally, we end this section with a few words about the word-suggestion feature of HySpell. The power of word-suggestion feature in any spell checker tool depends on the number of steps that is required for a misspelled word to transform into a legal or correct word, with respect to the lexicon-base. This means that if a misspelled word has three or more characters in the word that are off from the intended correct spelling, the word-suggestion feature becomes less accurate. In this case, the best advice to the end-user is to correct one or two characters of the given word manually and reattempt the proof reading.

Difference between HySpell, hyspell and hunspell

The development of HySpell would have taken much longer if it was not for the open-source spell-checker engine called hunspell. This later application however, needed porting and extensions to the .NET platform in order for it to be integrated with Microsoft Office 2007. The engine part of the HySpell is kept open-source in accordance to the GPL license agreement on the original hunspell base source. This base source code is available on the Internet at several sites.

Our name HySpell is actually non-brainer, it is derived from the HY language identifier for Armenian in the UNICODE standard and the English verb "Spell". This name is a little regretful, as I later found

that there existed an old Armenian spell checker called `hy_espell` (i.e. combination of “`hye`” and “`spell`”). So, it is important to know that our HySpell Armenian spell checker software that is posted at our site HySpell.com is different from the old `hy_espell`, and has nothing to do with Bytec Computers.

Some of the advantages of our Armenian spell checker include the following:

- (1) HySpell is tightly integrated inside Microsoft Office 2007 Word, and therefore very natural to use when writing a document in Word.
- (2) HySpell exposes the entire lexicon-base along with the affix rule files, so that the user may further customize to suit their special linguistic needs.
- (3) HySpell’s entire source code is also available as a separate product. Therefore, the product is considered open-source (under certain legal restrictions, see the **License.txt** file for more details about the usage terms, or contact support@hyspell.com for more information).
- (4) HySpell’s lexicon-base contains around 156,000 Armenian root words, and is the largest such lexicon on the market. The lexicon-base is supplemented with more than 300 grammar rules, spanning over 5 million Armenian legal inflections and words.
- (5) HySpell is fast and practical to use in Microsoft Office 2007 environment.
- (6) Finally, the current release of HySpell is a preliminary version 1.0, and therefore, subsequent releases of HySpell will be even more accurate and will include Armenian Thesaurus and Armenian-English electronic dictionary tightly integrated inside Microsoft Office products.