# Hunspell-fi in Kesäkoodi 2006: Final report

Harri Pitkänen

3rd September 2006

# Contents

**Introduction** **2**

**History** **3**

**Project plan** **5**

**Project results** **6**
- Spelling suggestions . . . . . 6
- Verb inflection tool . . . . . 7
- Vocabulary management tool Joukahainen . . . . . 8
- Other results . . . . . 10
- Problems . . . . . 10

**Future projects** **12**

**Appendix 1: Original project plan** **13**

**Appendix 2: License** **15**

# Introduction

This document is written as a final report of my "Kesäkoodi 2006" (KK2006) project sponsored by COSS and its supporting organisations. Through KK2006 five students were given a possibility to work on a free / open source (FOSS) software project of their choice as detailed in their project plans. My project was in the field of Finnish spell checking and technology to support development of collaboratively edited spell checker dictionaries for languages that have relatively complex morphologies.

This report starts with a look back to the history of the project and status of FOSS linguistic software from the point of view of a Finnish user. I believe this to be important considering that our project is much less known than most of the other projects in KK2006. In the light of our history I will introduce the goals of my work in KK2006 and what was actually achieved during the three summer months. Finally I will present a brief look to the future in the form of a list of interesting project that could be worth pursuing.

# History

My work on Finnish spell checking started in spring 2005 when I successfully managed to port the binary only but freely distributable spellchecker and hyphenator Soikko to work with OpenOffice.org version 2.0. Soikko was developed by Pasi Ryhänen, and its history is not very well known publicly. As far as I know, Pasi started writing Soikko in 1990's as a tool to help solve crossword puzzles. Soikko became a very advanced tool, but despite of many request it was never released as free software and development seems to have been abandoned completely few years ago. Pasi also wrote the initial integration code for OpenOffice.org 1.0 and 1.1. That code was luckily released under the LGPG/SISSL combination because it used some of the lingucomponent code from OpenOffice.org. The existence of that code under a free license allowed me to integrate Soikko with newer OpenOffice.org without any further knowledge of the binary only component itself.

I originally intended my work with Soikko integration be an one time contribution, but things did not go as I planned. The code needed fixing, and soon I started thinking about what would be needed to get a decent free replacement for Soikko. I had not managed to actually do anything before I read about Hunspell, a free spellchecker written to satisfy the needs of Hungarian language. Hunspell was an evolutionary improvement over Myspell (used at the time in OpenOffice.org), adding among other thing second level of affix stripping and some limited support for compound words, but otherwise retaining the original design of Myspell. This seemed ideal to me, as it fixed the worst problems that made Myspell unusable for Finnish. I started to write an experimental affix file (used in Hunspell to describe the morphology of a language) with very little knowledge of what I was supposed to do. For basic noun inflection things started quickly to look very good: it was not difficult to exceed the quality of old Myspell implementation, but I had to exclude clitics, compounding, derivation and verbs from the initial implementation.

By September 2005 few other people had joined my effort, and they started to collect and classify nouns for our vocabulary, which by the end of the year had reached the size of some 5000 words. In October I obtained the domain name hunspell-fi.org and moved the project there as it started to become unsuitably complicated to host in my personal web space at University of Jyväskylä. We also set up a mailing list and Reijo Tomperi created a vocabulary collecting application (still in production use) for suggesting missing words for our vocabulary. Unfortunately my studies at that time did not fit well with this hobby project. They contained no linguistics but a mixture of physics, mathematics and computer science, the weight being on using computer simulations to find solutions for certain engineering problems in quantum information theory.

This started to become a real problem as I could not develop the affix file and other necessary things that were needed by the people collecting words to our vocabulary. Our progress started to slow down.

Early this year I was considering different options to improve the situation. In February I was pointed to the possibility for applying KK2006 funding, which seemed to be the most effective way to go forward. I wrote my initial application, but only a week after submitting it Hannu Väisänen introduced his own implementation of Finnish morphology written in Malaga. This "Suomi-malaga" was written primarily for text indexing and therefore was intentionally designed to accept common misspellings and had very free compounding rules. But all of these problems seemed correctable, and at that time I estimated that improving Hunspell and our vocabulary to match the features of Suomi-malaga would have required at least six months of additional work. Indeed, my original application for KK2006 estimated that the 1.0 release of Finnish Hunspell dictionary, affix files and the fixes needed for Hunspell to support them would have been ready during the first half of 2007. Even with me being able to work on it full time during the summer it would have been hard to achieve that goal.

# Project plan

For the finals of KK2006 selection process I had the possibility to submit a refined proposal. From that proposal I removed the parts that concerned developing derivation and compounding rules for Hunspell (the selection committee had recommended me to decrease the number of non-programming tasks anyway which made my decision easier) and replaced that with a plan to create spelling suggestion code for a Malaga based implementation. The majority of my plan had consisted of a web application and database for managing the spell checker vocabulary; that part of the plan remained unchanged as the need for such application was the same for the Malaga based spell checker and the Hunspell based one.

The short version of my final project plan was as follows:

- Write an algorithm to create suggestions for incorrectly written words.
- Write code to list characteristic inflections for verbs according to their inflection class.
- Write a web application for distributed development of the vocabulary. It should be able to
  - add, edit and remove words and associated meta information,
  - help in manual proofreading of existing vocabulary,
  - automatically detect and avoid certain types of errors in the data,
  - output the spell checking vocabulary and other types of data as needed,
  - limit editing rights to registered users and
  - be extensible if there is need to use the data for other purposes, or use the program for languages other than Finnish.

The original, more detailed plan is available in Appendix 1. I will return to these points in the next chapter, where I will describe what exactly was done and which of the features turned out to be difficult or less important to implement.

To make sure that my work would spread evenly through the summer I made an additional work plan that contained target dates for each of the features in my official work plan. This plan was made available at http://www.hunspell-fi.org/kesakoodi.html and it contained also items that were not in my KK2006 plan.

# Project results

## Spelling suggestions

Spelling suggestions were the first part of my project, and I had estimated that two weeks would be needed to design and implement the needed algorithms. I started by writing a prototype algorithm in Python, and once the results were good enough I reimplemented the algorithm in C.

The design goal of the algorithm was simple: it was supposed to come up with the most probable correction for a misspelled word within the constraint that CPU time used on a 1 GHz x86 processor should on average remain near 0.1 seconds. This requirement comes from the fact that in typical user interface the suggestions are generated on demand when user opens a pop-up menu on an incorrectly spelled word, and a pop-up menu is generally expected to open instantly with no noticeable delay. The internal implementation of Malaga does not allow directly searching the vocabulary, so the only way to find the suggestions is to create variations of the input string and test whether they are valid words or not. In practise the aforementioned CPU time constraint limits the number of strings to be tested in an average case to about 300.

The basic operations that the suggestion algorithm should perform on an input string are single character deletions, single character insertions, single character replacements and two character swaps. Assuming that we have a string of $n$ characters from an alphabet consisting of $m$ characters the total number of these operations is

$$\begin{aligned} N_{tot} &= N_{del} + N_{ins} + N_{repl} + N_{swap} \\ &= n + (n+1)m + n(m-1) + \frac{n(n-1)}{2} \\ &= \frac{n^2}{2} + n(2m - \frac{1}{2}) + m \end{aligned}$$

We assume here that degenerate suggestions resulting from swapping two identical letters do not have significant effect. Even in the optimistic scenario, where we limit our alphabet to consist of the common Finnish characters "abdefghijklmnoprstuvyäö" we have $m = 23$ and must be able to properly handle words of at least ten characters, that is $n = 10$. This will lead to $N_{tot} = 528$ which is already slightly too much. However we currently have the need to create suggestions from an alphabet having $m = 33$ and realistically we also need to have at least $n = 20$ which gets us to $N_{tot} = 1543$. This would be totally unacceptable. There is also need to suggest splitting word to two parts, adding hyphens and

replacing simultaneously more than one front vowel with corresponding back vowel (or back vowel with front vowel).

Considering these limitations I adopted the following algorithm:

1. Create suggestion candidates in the order of their "closeness" to the input string. This closeness was not defined using any existing metric but by ordering different insertions, deletions, replacements, swaps and special operations by their probability using my own typos as an informal guide. For example, I took into account the distance of the keys on the keyboard and the fact that certain letters are pronounced similarly and could therefore be mixed in the written text as well.

2. These suggestions are tested and acceptable Finnish words are stored in an array. The testing is continued until we either have reached the CPU limit or have produced 15 acceptable words.

3. The accepted suggestions are sorted using a stable sort by the number of allomorphs in the word. This ensures that simple words are given higher priority than compounds and prefixed words, which in many cases are formally valid nonsense. Stable sort ensures that the original order by closeness to the input string is preserved for words with equal number of allomorphs.

4. The list is truncated to five most probable suggestions and returned to the calling application.

This algorithm seems to be giving rather good results, at least with real word input. It does not work so well if words contain artificial, random changes. Although the algorithm contains elements that are based on subjective evaluation on what kind of typos are the most important to find, it was improved during the summer based on feedback from testers and should be rather usable.

From the original project plan the previous algorithm implements everything except taking into account the frequency of different inflections. Implementing this would have required changes to the communication protocol between Suomi-malaga and the spell checking library libvoikko. I concluded that implementing such change would have required too much effort to do separately. It will hopefully be done in the near future among the other changes that are needed to provide a way for libvoikko to take advantage of the vocabulary information within Suomi-malaga.

## Verb inflection tool

In practise verb inflection tool was done as a part of the vocabulary management tool, but because it was a separate item in my project plan I will describe it here separately as well. It turned out to be the simplest thing in my project, as implementing this on top of my previous work on noun inflection only took a couple of days, most of which was used to analyse the inflection classes of the verbs and coding the actual inflection data for different classes. Apart from the characteristic inflections that need to be known to fully determine the inflection class for a verb I added the possibility to display derivations as well. This was needed because in many cases the "classes" in Suomi-malaga differ only by the

derivations: in these cases looking only at the inflections is not enough to find the correct class.

## Vocabulary management tool Joukahainen

Joukahainen was the main product of my KK2006 project. Apart from the first two weeks I worked on it through all of the summer. I had planned my work in periods of one, two or three weeks and during these I first tried to achieve the goal of the current period and then used the extra time to work on other aspects of Voikko. This way I managed to stay on the schedule quite well, but I intentionally did not want to proceed too fast either.

The goals and in which order I achieved them have been documented in my project blog and planning page. I will therefore not repeat them here, but will instead describe my work from the point of view of my original project plan.

- Adding, editing and removing words and associated meta information. Apart from removing words, all of these have been implemented. I concluded that ability to remove words would do more harm than good. The situation is similar to that in a popular bug tracking software Bugzilla: removing bugs would destroy the potentially valuable information they may contain. If incorrect words could be removed from Joukahainen, the reason for removal would no longer be easy to find and someone could sooner or later try to add the same word back. In order to work around this issue I decided to implement two flags, "Incorrect word" and "Moved" where the latter means that word is correct but not in the word class where it was originally stored. An alternative for these flags would be to implement a word status field and a possibility to move words from one class to another. Word status field is likely to be implemented later, but moving is a bit harder: there is again a possibility to lose data because of different set of meta information fields available for different word classes. All of the meta information fields in the original project plan were implemented, but many others were invented and added through the summer. We now have a possibility to store two different inflection classes (current and historical) and several flags to control the use of word in compounds, limit the derivation or describe the style or field of usage of the word. Finally we can also give a short description for the word and add links to a Wiktionary, so that Joukahainen could quite easily be used as a replacement for ordinary dictionary.

- Manual proofreading of existing vocabulary. This was done by implementing a concept of "vocabulary wide review tasks". A task is defined by a SQL query returning a subset of words in the database and a description of the task. Any registered user can request set of words from a task to work on. The user may then mark the words as reviewed, or pass them on for later review. This concept is more general than the one proposed in my project plan. It can be used for checking the inflection classes (as the characteristic inflection for the word is displayed in the word editor) but also any other aspects of the word.
  The separation of privileges (vocabulary reviewers not having rights to edit the words) was not implemented. Instead, the logging mechanism in

Joukahainen was made more comprehensive. It keeps track of all changes to the data and therefore the risk in allowing even less experienced users to edit the data is rather low.
As I write this, the first test run of verb classification check has been running for two weeks. It is now 45 % complete (that is, about 2500 verbs have been checked) and about 20 possibly problematic words have been identified.

- Automatically detecting and avoiding errors in the data. The inflection code and Joukahainen to Malaga conversion utility have been set to enforce certain validity restrictions for the words. These have mostly been implemented by requiring that words in given inflection class match certain regular expression. Another efficient way to prevent errors is to avoid storing redundant data. In the original Suomi-malaga lexicon database there are fields that can almost always be derived from the other fields, and while importing data to Joukahainen I managed to find several consistency errors that translated to actual bug fixes to the vocabulary. In fact, almost one percent of the words contained such errors, and in most of the cases the errors could have caused incorrect spelling suggestions and other issues for the users of Voikko.
  For compound words such errors can be avoided by associating the compounds with the trailing part of the word, which will ensure that they will automatically be inflected and derived correctly.

- Data output functionality. In the final design of Joukahainen, data output has been implemented as a feature of word searching functionality. This means that the user may retrieve parts of the vocabulary by entering suitable search criteria in the search form and requesting the output either as ordinary html search results or in the form of Suomi-malaga lexicon file. Other output formats can be added quite easily as needed, and the output formatters are free to access the database to retrieve any related data they need.

- Limiting edit rights to registered users. This was implemented as planned. Registration is required to do any changes to the data, but unregistered users can browse and query the database rather freely. There are only two types of users, administrators and non-administrator users. The only additional right administrators have is to add new users. No additional levels of access rights are currently planned, because logging of changes ensures that we will always be able to revert unwanted changes should there ever be need to do so.

- Extensibility and use for languages other than Finnish. This was a part of the project where I spent more effort than I had originally planned. The extensibility of Joukahainen has been ensured in the following ways:

  - The database schema does not need to be changed in order to add or remove meta data fields or word classes. It is sufficient to add entries to a special database table holding information about attribute names, types and the word classes that the attribute is used in. In case of attributes holding text values the page template for word

editor needs to be changed as well. It is possible to perform these changes even while the application is in production use.

- All user visible strings in the program code can be localised using the gettext tools commonly used in free software localisation.
- All language specific material is separated and available as language packs. These can include information about the vocabulary meta data fields mentioned above, program localisation files, page templates and language specific program modules.
- Other parts of the program code have been designed for easy extensibility where possible. For example the word editor consist of language specific page template and editor components that can be added as needed. This will make it easy to add support for number entry fields, if we ever need to store numerical information in the database (quite likely to happen sooner or later).

Outside my project plan I exchanged a few emails with Kevin Scannell who has expressed interest in using Joukahainen for creating Hunspell dictionaries for languages that do not yet have them. Based on feedback from him I extended the word entry functionality to allow the use of raw candidate words collected by a web crawler. This list of candidate words may also contain some classification information if something like that is available. It is uncertain if this functionality will be used for Finnish vocabulary development or not. It could with relatively little extra effort replace the previous word collecting application from Reijo Tomperi, but at this point I do not consider that to be a priority because the old system works quite well.

## Other results

The planned projects as whole turned out to take somewhat less time than I had prepared for. This allowed me to work on other aspects of Voikko spell checking system that I had not included in my plan. I spent some time to improve tmispell-voikko (an ispell wrapper allows programs designed to use ispell take advantage of Voikko as well) and Oo2-voikko (OpenOffice.org spellchecker and hyphenator extension). I also managed to fix language correctness issues in Suomi-malaga. This was largely made possible by the active testers on our mailing list, as I did not have much time for testing of the linguistic correctness myself.

I had anticipated that the release of Voikko 1.0 on August 13th (the date had been set well in advance) would largely represent the state of the system it was in June. This turned out to be quite pessimistic estimate. We were able to make significant improvements even while I was primarily working on Joukahainen.

## Problems

There were also areas where I would have wanted to make more progress. The most important was work with the inflection classification system. It turned out

to require a lot of work to move some of the more obscure words from Suomimalaga to Joukahainen. I had not reserved much time for this type of work, and therefore my goal of getting Joukahainen to full production use by the release of Voikko 1.0 could not be reached. I believe that having an easy to use and logical classification is required for Joukahainen to fulfil its goals. Currently the conversion works for the majority of the words, but because natural languages tend to have a lot of exceptions it seems that the principle of last 10 % of the work taking 90 % of the time unfortunately applies here. I still hope to reach the level of 99 % completeness during this September, which should be enough to declare Joukahainen ready for full production use.

# Future projects

Free software linguistic tools for Finnish (and other languages) still need a lot of work. In no particular order of importance here are some that interest me personally:

- Examining the theoretical foundations of out current implementation and comparing it to alternative implementations. Formalising the morphological grammar could help building more efficient implementations, as the theory of formal languages and parsers has been researched quite a lot during the last few decades. From the current implementation of Suomi-malaga we can only say that it accepts a recursive language, but that does not help much when we want to optimise the implementation. In order to use the more powerful tools it seems that we should formulate the morphological rules in such way that the language would be at least context free, maybe even regular. Whether this means that we should continue to use Malaga, switch to something else or write our own parser from scratch remains to be seen.
- Move the the information about morphological rules (inflections in different classes, word combining and derivations) to an application independent form. Joukahainen does this for the vocabulary, but why stop there?
- Facilitate using more information from the parser backend (Suomi-malaga) in libvoikko to make better decisions about word correctness or spelling suggestions. This should be almost trivial if the previous two projects were properly implemented.
- Create Finnish thesaurus. This could be done by adding support for inter word relations to Joukahainen.
- Finnish grammar checking. Very difficult subject, but something usable could be done by simply listing some common erroneous phrases or word combinations. The Google Summer of Code 2006 project by Bruno Sant'Anna concentrated on grammar checker API for OpenOffice.org. Oo2-voikko could be extended to use this to do some interesting things.
- Better integration of existing tools in free applications. A lot could be done by using Enchant instead of some specific spellchecker backend.

# Appendix 1: Original project plan

The following was my original project plan (in Finnish):

- Algoritmin toteuttaminen väärin kirjoitettujen sanojen korjausehdotusten tuottamiseksi. Malaga ei tarjoa tällaista valmiiksi, ja Hunspellin vastaava algoritmi ei huomioi suomen kielen erityispiirteitä. Kun väärin kirjoitetulle sanalle etsitään kandidaatteja oikeiksi muodoiksi, on huomioitava se, että tietyt kirjaimet sekoittuvat helpommin keskenään kuin jotkin muut. Esimerkiksi näppäimistöllä lähekkäin olevat merkit menevät helposti virhepainallusten takia sekaisin, samoin kuin äänteinä samankaltaiset kirjaimet (esimerkiksi a ja ä). Lisäksi korjausehdotuksissa yhdyssanoille ja harvinaisille taivutusmuodoille kannattaa antaa pienempi prioriteetti (tätä eivät olemassa olevat algoritmit huomioi lainkaan). Algoritmia pitää myös yrittää optimoida siten, että se löytää hyviä korjausehdotuksia korkeintaan muutaman sadan yrityksen jälkeen.

- Ohjelma verbien automaattiseen taivuttamiseen tiettyihin "karakteristisiin" taivutusmuotoihin sanaston luokittelun avuksi. Tarkoitus on siis toteuttaa rajoitetusti oikoluvulle käänteinen prosessi jossa ohjelma listaa verbille tietyt taivutusmuodot, kun sille syötetään perusmuoto ja taivutusluokitus. Tätä ei voi tehdä olemassa olevilla ohjelmilla, mutta tämä on kuitenkin tärkeä apuväline sanastoa kerääville ja tarkistaville henkilöille. Pelkkä sana ja sen taivutusluokka yksinään eivät kerro juuri mitään muille kuin koneelle. Taivutuksia voi toki miettiä päässä, mutta tämä on osoittautunut hitaaksi ja virheitä tulee liikaa. Helpompi on vain tarkistaa, että koneen listaamat taivutukset ovat oikein. Tällaisen ohjelman olen jo tehnyt nomineille, mutta ohjelmaa pitäisi kehittää edelleen jotta se toimisi myös verbien kanssa.

- WWW-pohjainen tietokantasovellus sanaston hajautettuun käsittelyyn. Tämä olisi kesäkoodiprojektin päätuote. Sovelluksen avulla pitäisi pystyä

  - Lisäämään, poistamaan ja muokkaamaan sanaston sanoja ja niihin liittyviä metatietoja. Metatietokentistä alkuvaiheessa on toteutettava ainakin sanaluokka, taivutusluokka ja astevaihteluluokka. Lisäksi tarvitaan vokaalisoinnun määritys niille sanoille, joille tavanomainen algoritmi ei toimi (monet vierasperäiset sanat) ja mahdollisuus merkitä sana kuuluvaksi jonkin erityisalan sanastoon tai puhekieleen.

- Oikolukemaan tehokkaasti valmista sanastoa. Tätä varten käyttäjille on tarjottava mahdollisuus saada sanan taivutukset nähtävilleen (käyttäen aikaisemmin tehtyä taivutusohjelmaa) ja antamaan arvionsa siitä, onko taivutus oikein. Samoin käyttäjän on voitava tuoda ilmi mahdolliset väärin kirjoitetut sanat. Kuitenkaan muutoksia ei oikoluvussa saa tehdä suoraan, vaan on varmistuttava siitä, että varsinaisia korjauksia saavat tehdä vain siihen oikeutetut henkilöt.
- Erilaisia tietojen järkevyystarkisteluja on pystyttävä tekemään helposti, ja tärkeimmät niistä on toteutettava heti ensivaiheessa. Esimerkiksi sanojen luokituksessa tietyt taivutusluokat voidaan automaattisesti sulkea pois sanan kirjoitusasun perusteella. Lisäksi sanastoon joudutaan jonkin verran lisäämään yhdyssanoja. Näiden kohdalla on varmistuttava, että taivutusluokitus vastaa aina sanan jälkiosan luokitusta, jos sellainen on erikseen sanastossa.
- Tekijänoikeuksien turvaamiseksi ja väärinkäytösten estämiseksi sovelluksen on vaadittava rekisteröityminen kaikilta käyttäjiltä.
- Sovellus pitää suunnitella siten, että sitä voidaan jatkossa laajentaa muihin tarkoituksiin, esimerkiksi synonyymisanaston rakentamiseen tai muiden kielien sanastojen kokoamiseen.
- Sovelluksesta on pystyttävä helposti saamaan ulos varsinainen sanasto, mutta tarvittaessa myös muitakin tietoja.

# Appendix 2: License

This document is licensed under “Creative Commons Attribution-ShareAlike 2.5”. The full license text is written below.

## Attribution-ShareAlike 2.5

CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS LICENSE DOES NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN "AS-IS" BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE INFORMATION PROVIDED, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM ITS USE.

### License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED. BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. **Definitions**

   (a) **"Collective Work"** means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this License.

   (b) **"Derivative Work"** means a work based upon the Work or upon the Work and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation,

or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work will not be considered a Derivative Work for the purpose of this License. For the avoidance of doubt, where the Work is a musical composition or sound recording, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered a Derivative Work for the purpose of this License.

(c) **"Licensor"** means the individual or entity that offers the Work under the terms of this License.

(d) **"Original Author"** means the individual or entity who created the Work.

(e) **"Work"** means the copyrightable work of authorship offered under the terms of this License.

(f) **"You"** means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

(g) **"License Elements"** means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, ShareAlike.

2. **Fair Use Rights.** Nothing in this license is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3. **License Grant.** Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

   (a) to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;

   (b) to create and reproduce Derivative Works;

   (c) to distribute copies or phonorecords of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works;

   (d) to distribute copies or phonorecords of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission Derivative Works.

   (e) For the avoidance of doubt, where the work is a musical composition:

      i. **Performance Royalties Under Blanket Licenses.** Licensor waives the exclusive right to collect, whether individually or via a performance rights society (e.g. ASCAP, BMI, SESAC), royalties for the public performance or public digital performance (e.g. webcast) of the Work.

ii. **Mechanical Rights and Statutory Royalties.** Licensor waives the exclusive right to collect, whether individually or via a music rights society or designated agent (e.g. Harry Fox Agency), royalties for any phonorecord You create from the Work ("cover version") and distribute, subject to the compulsory license created by 17 USC Section 115 of the US Copyright Act (or the equivalent in other jurisdictions).

(f) **Webcasting Rights and Statutory Royalties.** For the avoidance of doubt, where the Work is a sound recording, Licensor waives the exclusive right to collect, whether individually or via a performance-rights society (e.g. SoundExchange), royalties for the public digital performance (e.g. webcast) of the Work, subject to the compulsory license created by 17 USC Section 114 of the US Copyright Act (or the equivalent in other jurisdictions).

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4. **Restrictions.** The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

(a) You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this License, and You must include a copy of, or the Uniform Resource Identifier for, this License with every copy or phonorecord of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this License or the recipients' exercise of the rights granted hereunder. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this License Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this License. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any credit as required by clause 4(c), as requested. If You create a Derivative Work, upon notice from any Licensor You must, to the extent practicable, remove from the Derivative Work any credit as required by clause 4(c), as requested.

(b) You may distribute, publicly display, publicly perform, or publicly digitally perform a Derivative Work only under the terms of this License, a later version of this License with the same License Elements as this License, or a Creative Commons iCommons license that

contains the same License Elements as this License (e.g. Attribution-ShareAlike 2.5 Japan). You must include a copy of, or the Uniform Resource Identifier for, this License or other license specified in the previous sentence with every copy or phonorecord of each Derivative Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Derivative Works that alter or restrict the terms of this License or the recipients' exercise of the rights granted hereunder, and You must keep intact all notices that refer to this License and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Derivative Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this License Agreement. The above applies to the Derivative Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Derivative Work itself to be made subject to the terms of this License.

(c) If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Derivative Works or Collective Works, You must keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or (ii) if the Original Author and/or Licensor designate another party or parties (e.g. a sponsor institute, publishing entity, journal) for attribution in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; the title of the Work if supplied; to the extent reasonably practicable, the Uniform Resource Identifier, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and in the case of a Derivative Work, a credit identifying the use of the Work in the Derivative Work (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Derivative Work or Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5. **Representations, Warranties and Disclaimer** UNLESS OTHERWISE AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE MATERIALS, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. **Limitation on Liability.** EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. **Termination**

   (a) This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Derivative Works or Collective Works from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

   (b) Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. **Miscellaneous**

   (a) Each time You distribute or publicly digitally perform the Work or a Collective Work, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

   (b) Each time You distribute or publicly digitally perform a Derivative Work, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.

   (c) If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

   (d) No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

   (e) This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified

here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.

Creative Commons is not a party to this License, and makes no warranty whatsoever in connection with the Work. Creative Commons will not be liable to You or any party on any legal theory for any damages whatsoever, including without limitation any general, special, incidental or consequential damages arising in connection to this license. Notwithstanding the foregoing two (2) sentences, if Creative Commons has expressly identified itself as the Licensor hereunder, it shall have all rights and obligations of Licensor. Except for the limited purpose of indicating to the public that the Work is licensed under the CCPL, neither party will use the trademark "Creative Commons" or any related trademark or logo of Creative Commons without the prior written consent of Creative Commons. Any permitted use will be in compliance with Creative Commons' then-current trademark usage guidelines, as may be published on its website or otherwise made available upon request from time to time. Creative Commons may be contacted at http://creativecommons.org/.