JavaSpecialists

Translating a Cretan Book using Java and ChatGPT API

Dr Heinz M. Kabutz

Last updated 2025-01-23

Vacation Time - Project Euler

◎ **Project Euler**

– 2023 was rough, all work no play

– End of December, finally time to relax

– Started solving maths puzzles with code

• www.projecteuler.net

• They get progressively more difficult

• Good thinking and coding practice

• Only catch: Up all night dreaming of better algorithms

– #21 in Greece at the moment

Looking Back Two Years

◉ **Constant tracking of body markers, such as**

- Sleep score
  - How well I slept
  - Total sleep
  - Sleep Efficiency
- Readiness score
  - How likely to be running a marathon that day
  - Heart Rate Variability

◉ **All tracked automatically with my Oura Ring**

- www.ouraring.com

Sleep Score

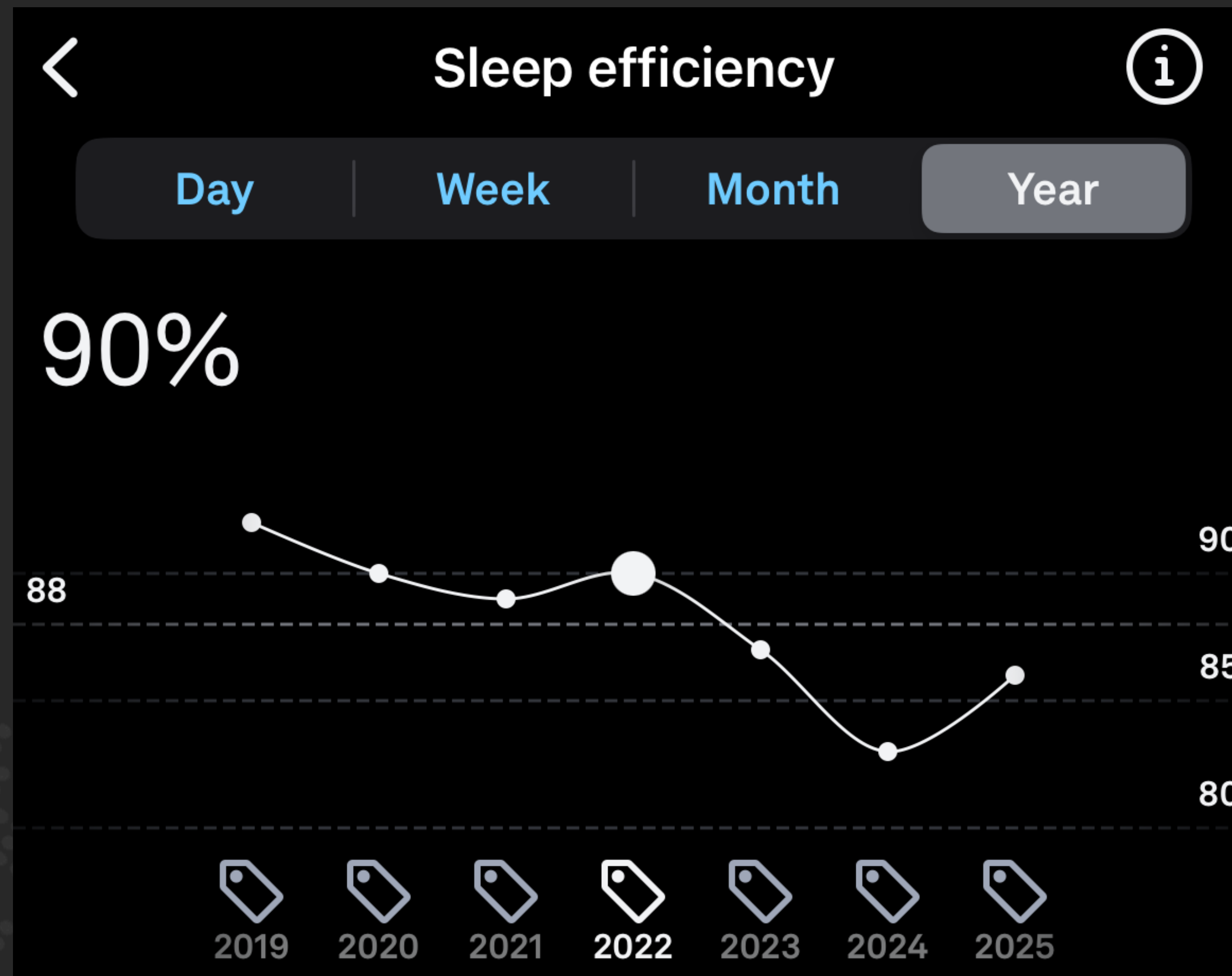◉ **Dipped down quite badly in 2023 and 2024**

Total Sleep

◉ **Dipped down quite badly in 2023 and 2024**

Sleep Efficiency

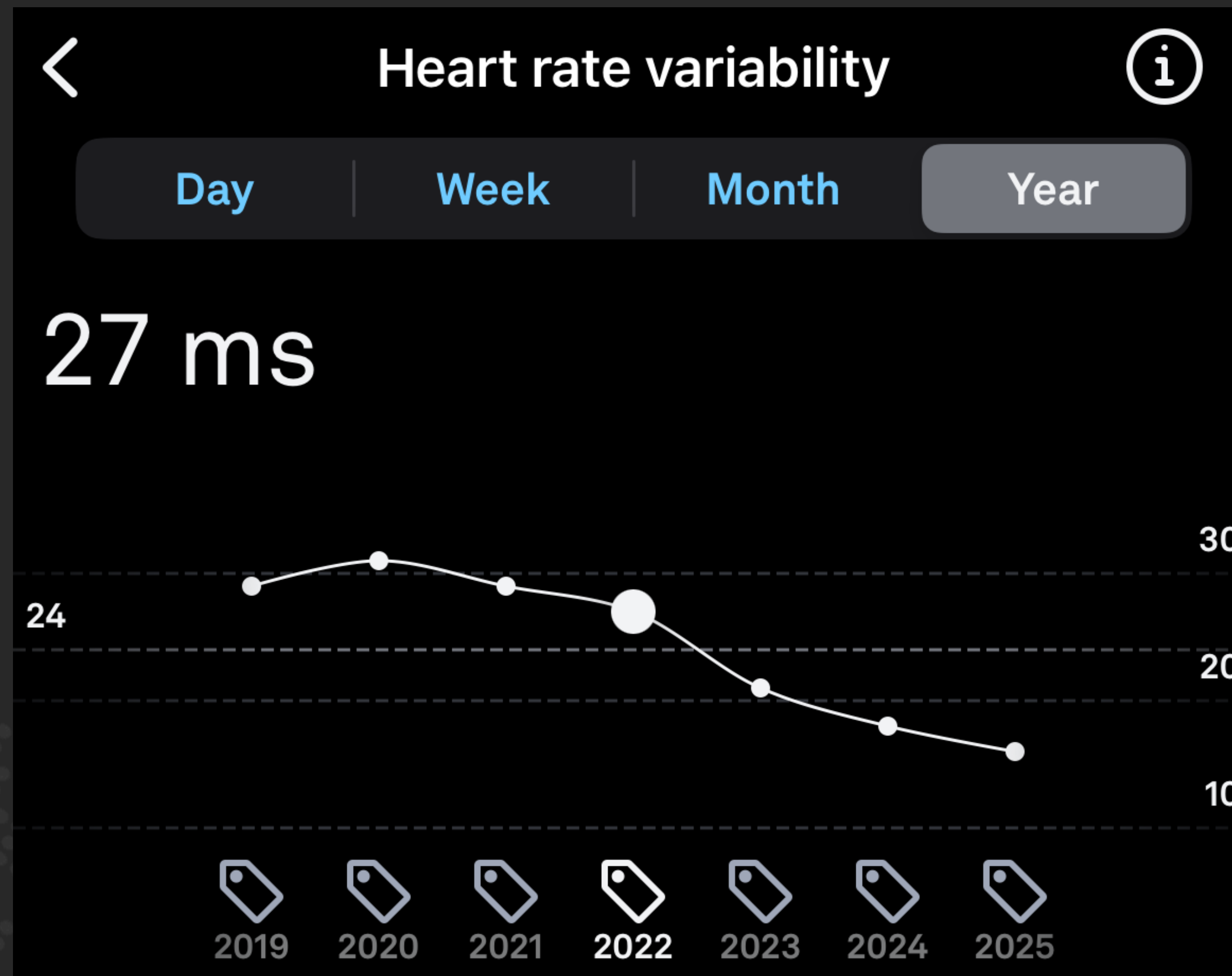◉ **We often use the excuse - "At least I slept well!"**

Readiness Score was also terrible

⊙ **No running marathons**

Heart Rate Variability

◉ **Higher is better - need to work on that for 2025**

Introducing Mr Manolis Koutrakis

◉ **Mr Manolis**

– 2nd oldest man in Chorafakia

- 91 years old

– Spent 20 years writing history book about our village and area

- But book is in Cretan Greek

Photo by Wouter Pardaens

Tsikoudia Kazani - Cauldron

◉ **Cretan "Medicine"**

– Mr Giannis "σμίξαμε"

– One week of working day and night

– Hundreds of liters of great tsikoudia

• Tastes better from cauldron

Photo by Wouter Pardaens

Photo by
Wouter Pardaens

The Book

◉ **My Village Chorafakia**

– Written in Cretan Dialect

– Beautiful flowery language

– Looked for someone to type it in

• 20 years of handwritten notes

• Almost 500 photographs

– Eventually the book was printed

• I bought it, couldn't understand it

• Asked if I could try translate it

ΤΟ ΧΩΡΙΟ ΜΟΥ
ΧΩΡΑΦΑΚΙΑ
ΑΚΡΩΤΗΡΙΟΥ ΧΑΝΙΩΝ

Photo by
Wouter Pardaens

First Attempt

◉ **WordPress Website with translation plugin**

– www.chorafakia.com

◉ **Translation automatically with Google Translate**

– Stored in database for later editing

◉ **Upgraded plugin and lost all the translations**

Next up - LibreOffice

◎ **Tried to convert the PDF to LibreOffice**

– Worked fine for things like footnotes of the Cretan dialect

– However, couldn't break up document into smaller chunks

• Divide and conquer

– Translation would require copy & pasting
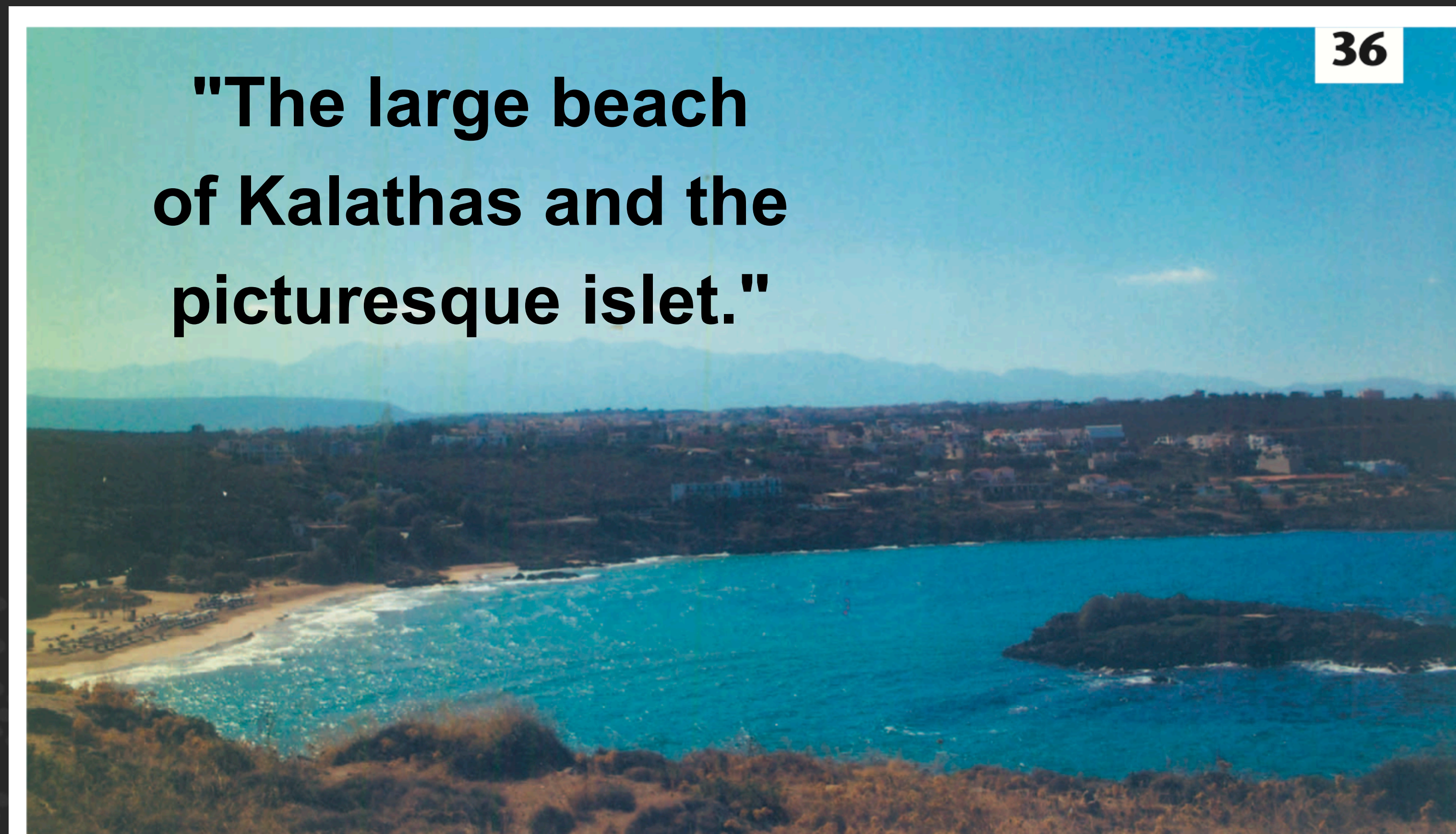
• Not practical at all

AsciiDoc

◉ **Next up: AsciiDoc**

– Text based markup format for writing documents and books

- Wrote Java Dynamic Proxies with this tool

- Not as powerful as LaTeX, but good enough

– Could break up text into hundreds of small chunks

– Could put everything in GitHub

– Produced PDF, EPUB, HTML5, etc.

Photos with Descriptions

◉ **Book has hundreds of photos from bygone eras**

– Each photo has a number and a description

"The large beach of Kalathas and the picturesque islet."

Η μεγάλη παραλία του Καλαθά και το γραφικό νησάκι.

Original Book Designed for Print

◉ **Not easy to read**

– Lexicon of Cretan words at the back

  • But you need to know what the Cretan words are

    – Sometimes marked with *

    – Lexicon words not sorted alphabetically

      » Duplicates, hard to find words

– All the colour pictures are at the end

  • Cheaper to print

  • Far more difficult to read

    – Constant flipping pages and searching for photo

Book Translation as Software Project

◎ **Software Engineering is related to informatics**

– Broke book into hundreds of small chunks

– Each chunk is in text format

• Easy to translate

• Stored in GitHub

– Book built using a Makefile and scripts

Enter ChatGPT 4.0

◉ **Google Translate won't cut it**

– Not terrible, but also not great

◉ **Signed up for ChatGPT 4.0**

– Incredible, even translated poetry

• Not rhyme, but definitely sounds poetic

Original: Poem to a Little Tree

Δεντράκι που είσαι στη σχισμή του βράχου ριζωμένο
πες μου που βρίσκεις την τροφή, που βρίσκεις το νερό
με ροζιασμένο τον κορμό και χιλιοπληγωμένο
πώς θρέφεις φύλλα πράσινα, πώς δένεις τον καρπό;

Κείνος που αστέρια γιόμισε ψηλά τον ουρανό,
κι έδωσε ξέχωρη φωνή στ' αηδόνι να λαλεί,
κι έδωσε γλύκα πλιότερη στης μάνας το φιλί
κείνος μου δίνει την τροφή, μου δίνει και νερό.

Google Translate: Poem to a Little Tree

Tree that you are in the cleft of the rock rooted
tell me where you find the food, where you find the water
with a rosy torso and thousands of wounds
how do you nourish green leaves, how do you bind the fruit?

He who stars shone high in the sky,
and gave a distinct voice to the nightingale to sing,
and gave the mother a sweeter kiss
he gives me food, he also gives me water.

ChatGPT 3.5: Poem to a Little Tree

Tree that you are in the crevice of the rock, rooted,

tell me where you find your food, where you find your water.

With your trunk rosy and heavily wounded,

how do you nurture green leaves, how do you bear fruit?

He who filled the sky high with stars,

and gave a distinct voice to the nightingale to sing,

and gave sweeter sweetness to a mother's kiss,

he gives me food, he gives me water too.

ChatGPT 4.0: Poem to a Little Tree

Little tree rooted in the crevice of a rock,

tell me, where do you find food, where do you find water?

With your reddened trunk and scarred a thousand times,

how do you nourish green leaves, how do you bear fruit?

He who filled the sky with stars high above,

who gave the nightingale its unique song,

who made a mother's kiss the sweetest,

he gives me sustenance, he gives me water.

ChatGPT 4o: Poem to a Little Tree

Little tree rooted in the crack of the rock,
tell me, where do you find food, where do you find water?
With your gnarled trunk, battered and scarred,
how do you nourish green leaves, how do you bear fruit?

The One who filled the heavens above with stars,
who gave the nightingale its distinct, beautiful voice,
who gave a mother's kiss its sweetest warmth,
He is the one who gives me food, who gives me water too.

DRY - Don't Repeat Yourself

◎ **Copied and pasted sections into ChatGPT prompt**

```
Translate into English, keeping AsciiDoc tags
intact:


.Πανοραμική φωτογραφία του λιβαδιού, όπως είναι
τσοι τωρινούς χρόνους.
[#Image_049]
image::../images/Image_049.png[]
```

– Produced this

```
.Panoramic photo of the meadow, as it is in
current times.
[#Image_049]
image::../images/Image_049.png[]
```

IntelliJ AI Integration

◉ **IntelliJ Ultimate has an AI integration**

– Let's try translating Image_049.adoc - demo 1

What about Generative AI?

◉ **Concern about**

- Copyrighted material

- Leaking sensitive business information, passwords, API keys, etc.

- Making all programmers unemployed
  - https://www.trueup.io/layoffs
  - And other jobs too
    - Models, authors, translators, artists, musicians, sex workers, etc.
    - Not hair dressers (yet)

Before we continue ...

◉ **A small gift for you - tinyurl.com/jcconf25**

   – Be sure to enable emails

Bulk Translation of Photo Blocks

◉ **Instead of one at a time, batch 50 photos**

  – Less latency for translation request, but kept on failing

  – ChatGPT has a limitation on number of tokens

tinyurl.com/jcconf25

Translating using the ChatGPT API

◉ **Instead of copy & paste, I asked ChatGPT how to use the API**

- "Show me how to send prompts to the ChatGPT API using Java"
  - ChatGPT showed me step by step how to do it with Apache HttpClient
- "Show me using the Java 11 HttpClient"
  - A minute later, changed the example to use the Java 11 HttpClient

tinyurl.com/jcconf25

ChatGPT API with HttpClient - Demo 2

```java
public class ChatGPTApiClientJava11 {
    private static final String API_URL =
            "https://api.openai.com/v1/chat/completions";
    private static final String API_KEY =
            "your_api_key_here";
            // Replace with your actual API key
    public static void main(String[] args) {
        sendPrompt("Tell me a joke");
    }
```

ChatGPT API using Java 11 HttpClient

```java
public static void sendPrompt(String prompt) {
    HttpClient client = HttpClient.newHttpClient();
    HttpRequest request = HttpRequest.newBuilder()
                .uri(URI.create(API_URL))
                .header("Authorization", "Bearer " + API_KEY)
                .header("Content-Type", "application/json")
                .POST(BodyPublishers.ofString("{\"model\": " +
                        "\"gpt-3.5-turbo\", \"messages\": " +
                        "[{\"role\": \"user\", \"content\": \"" +
                        prompt + "\"}]}"))
                .build();
```

ChatGPT API using Java 11 HttpClient

```java
        try {
            HttpResponse<String> response = client.send(
                    request, BodyHandlers.ofString());
            System.out.println(response.body());
        } catch (IOException | InterruptedException e) {
            e.printStackTrace();
        }
    }
}
```

Json using GSON, etc.

◉ **We should use a library for Json such as GSON**

– Again, we can ask ChatGPT to create the classes

```java
record ChatGPTRequest(String model, List<Message> messages) { }
record Message(String role, String content) {
    public Message(String content) {
        this("user", content);
    }
}
record ChatResponse(String id, String object, long created,
                    String model, List<Choice> choices, Usage usage,
                    String system_fingerprint) { }
record ErrorResponse(String message, String tokens, String code) { }
record Choice(int index, Message message, String finish_reason) { }
record Usage(int prompt_tokens, int completion_tokens, int total_tokens) {
```

Translating all the AsciiDoc files

◉ **We can translate each file sequentially like this**

```java
try (Stream<Path> stream = Files.walk(start)) {
    stream.filter(Files::isRegularFile)
        .filter(path -> !path.toString().contains("locations/"))
        .filter(path -> path.toString().endsWith(".adoc"))
        .map(Path::toUri)
        .map(URI::toString)
        .forEach(srcFilename -> translate(srcFilename, language));
}
```

– Why would it be a mistake to convert this to a parallel stream?

How long does the translation take?

◉ **Models have different speed and accuracy**

– Number of request and response tokens

– Cost of each token

– Rate limits on how many tokens per minute etc.

Model	Cost / 1M input tokens	Cost / 1M output tokens
Gpt-4o-mini	$0.15	$0.60
gpt-3.5-turbo-0125	$0.50	$1.50
gpt-4o	$2.50	$10
gpt-o1-mini	$3	$12
gpt-4-1106-	$10	$30
gpt-o1-preview	$15	$60

Structured concurrency virtual threads

◉ **Use a virtual thread executor service for the HttpClient**

```java
private static final HttpClient client = HttpClient.newBuilder()
            .executor(Executors.newVirtualThreadPerTaskExecutor())
            .build();
```

◉ **Fork each of the translation tasks**

```java
try (var pathStream = Files.walk(startPath)) {
    try (var scope = new StructuredTaskScope.ShutdownOnFailure()) {
        pathStream.filter(Files::isRegularFile)
                .filter(path -> !path.toString().contains("locations/"))
                .filter(path -> path.toString().endsWith(".adoc"))
                .map(Path::toUri)
                .map(URI::toString)
                .forEach(srcFilename -> scope.fork(() -> translate(srcFilename, language)));
        scope.joinUntil(until).throwIfFailed();
    }
}
```

Let's try Afrikaans - Demo 3

◉ **We want to see how fast and how much**

– We will monitor how much it costs to use the ChatGPT API

- https://platform.openai.com/account/billing/overview

– First with a Semaphore of 1 and gpt-3.5-turbo-0125

- Making it single-threaded

– Next we increase Semaphore to 100

– Then with model gpt-4-1106-preview and gpt-4o

Copying Symbolic Links

◉ **The directory structure has a lot of symbolic links**

– Images, locations, etc.

◉ **Asked ChatGPT to write the code for me**

– See MagicBookTranslator.copySymbolicLinks(source, target)

– Used walkFileTree() with a FileVisitor

• Email from a famous Java Champion:

– "You had a hunch that the code from your chatbot might not have been the best. The Files.walkFileTree method is needlessly complex. Just use Files.walk."

– Let's try do it with Files.walk() ... see copySymbolicLinksStream() demo 4

LangChain4j

◉ **Becoming the defacto standard for Java and AI**

– Standardized access to multiple LLMs and vector stores

– Tools for prompt management, chat memory, output parsing

– Integration with Java frameworks

– Lots of examples and documentation

– Active Development with engaged community

◉ **Let's ask it for a joke - demo 5**

What about the other code?

◉ **We can look at it, but ...**

– ChatGPT wrote most of the code

   • You can just ask it to do the same for you

Where to next?

◉ **Generative AI is already quite good**

  – However, a lot of companies are reluctant

  • Who owns our ideas, where is the generated code coming from?

  – New York Times is suing Microsoft for copyright violation

  – Oracle sued Google for copying the Java API for Android (but lost)

  • IntelliJ AI uses OpenAI ChatGPT

  – Doesn't use our prompts for training

  – They offer on-prem LLM with their enterprise option

  • Companies like Cohere.com offer LLM licenses

  • Microsoft Azure can also partition our data better than OpenAI

Whilst we are taking questions ...

◎ **Remember to get this - make sure to enable emails!**

– https://javaspecialists.teachable.com/p/data-structures?
coupon_code=JCHAMPIONSCONF2025 or tinyurl.com/jcconf25